2014

Data-Driven Implementation To Filter Fraudulent Medicaid Applications

Muhammad Nader Suleiman
North Carolina Agricultural and Technical State University

Data-Driven Implementation to Filter Fraudulent Medicaid Applications

Muhammad Nader Suleiman

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Computer Systems Technology

Major: Information Technology

Major Professor: Dr. Rajeev Agrawal

Greensboro, North Carolina

2014

The Graduate School
North Carolina Agricultural and Technical State University
This is to certify that the Master's Thesis of

Muhammad Nader Suleiman

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2014

Approved by:

Dr. Rajeev Agrawal Major Professor	Dr. Ibraheem Kateeb Committee Member
Dr. Cameron Seay Committee Member	Dr. Clay Gloster Jr. Department Chair

Dr. Sanjiv Sarin
Dean, The Graduate School

Biographical Sketch

Mr. Muhammad Nader Suleiman, a resident of Greensboro, North Carolina, is an IT professional whose work experience includes troubleshooting and database administration in a fast-paced environment. A dedicated worker with strong analytical and communication skills, he has a strong desire to fulfill his responsibilities and achieve outstanding results. In 2006, he earned his Bachelor of Science degree in information systems from the University of North Carolina at Greensboro. In 2013, he joined the information technology graduate program in computer system technology at North Carolina Agricultural and Technical State University. While pursuing his degree, he worked as a graduate research assistant for the NCAT Department of Human and Health Development. He also was a member of NCAT Graduate Student Advisory Council. He has published research at the 2013 International Conference on Management of Emergent Digital Ecosystems in Luxembourg and the 2014 IEEE Southeast Conference. He also presented research on data-driven approach to preventing Medicaid fraud at the Ronald E. McNair 28th Annual Celebration and 12th Annual Research Symposium and published research on dynamic disease forecast networks using family medical history at the Health care Informatics 2013 IEEE International Conference.

Dedication

I dedicate my thesis to my family. A special feeling of gratitude to my loving parents, Professor Dr. Nader and Raeda Suleiman, who provided me with caring support, words of encouragement, blessing, and financial support in my entire pursuit for higher education. My beloved wife Sawsan and wonderful daughters: Maryam & Yara for their love and being there for me throughout my graduate program. My brother Heitham, who was the constant source of support and contribution to my family throughout my thesis work challenges. My Sister Maha and my brother-in-law Nabil Warad, for caring my family and me under their wings during serve winter storms and difficult times. Last but not least, my uncles Kamel and Kamal, for their continuous encouragement, motivation, and inspiration during the pursuit and composition of my graduate degree.

Acknowledgements

Foremost, all praises to Allah, the Most Gracious and the Most Merciful, for giving me the strengths, ability, and blessing in completing this study. Second, my deepest appreciations to my committee chair Professor Dr. Rajeev Agrawal, for his supervision and constant support. His patience, motivation, and immense knowledge throughout the experimental and thesis study have contributed to the success of this research. His timely and resourceful contribution helped me all the time of research and writing of this thesis to shape it in the final form.

I would like to thank my Professor Dr. Cameron Seay, whose work demonstrated to me that concern for information technology supported by enterprise and cloud computing in comparative modern technology environments, should always transcend academia and provide a mission for our future. His invaluable help of constructive comments and advice throughout the thesis implementation phase have contributed to the success of this research presentation.

In addition, my sincere thanks to the department chair Dr. Clay Gloster and Professor Dr. Ibraheem Kateeb, for their encouragements, insightful comments, support and feedback, and the willing to always help all throughout my graduate program.

I would like to thank my good friend Dominic Mensah, who was pushing me to new involvements and always willing to help during my study. It would have been a lonely experience without him. Many thanks to fellow thesis mates Yolanda Baker, Curtis Jackson, and Jeffrey Anu for their guidance, suggestions, and helping me polish my defense presentation.

Last but not least, my deepest gratitude goes to my wife, Sawsan Suleiman. I'm forever indebted to her for being understanding, patient, and all her dream sacrifices. Without Allah and everyone's persistent help, this dissertation would not have been possible.

Table of Contents

List of Figures

List of Tables

Abstract

There has been much work to improve IT systems for managing and maintaining health records. The U.S government is trying to integrate different types of health care data for providers and patients. Health care fraud detection research has focused on claims by providers, physicians, hospitals, and other medical service providers to detect fraudulent billing, abuse, and waste. Data-mining techniques have been used to detect patterns in health care fraud and reduce the amount of waste and abuse in the health care system. However, less attention has been paid to implementing a system to detect fraudulent applications, specifically for Medicaid. In this study, a data-driven system using layered architecture to filter fraudulent applications for Medicaid was proposed. The Medicaid Eligibility Application System utilizes a set of public and private databases that contain individual asset records. These asset records are used to determine the Medicaid eligibility of applicants using a scoring model integrated with a threshold algorithm. The findings indicated that by using the proposed data-driven approach, the state Medicaid agency could filter fraudulent Medicaid applications and save over $4 million in Medicaid expenditures.

CHAPTER 1

Introduction

1.1 Background

The U.S. health care system has two federal health programs: Medicare and Medicaid. Medicare is a federal program that provides health insurance coverage for individuals aged 65 or older or individuals under age 65 with certain disabilities or conditions such as end-stage renal disease. Medicare has four parts:

A. Hospital insurance—provides payments to cover inpatient care in hospitals, including critical hospital services, skilled nursing facilities, and some home health care.
B. Medical insurance—provides payments to cover hospital outpatients, including doctors' services, preventive services, physical and occupational therapists, some home health care, and medical equipment.
C. Health plan coverage—provides payment to health plans to cover services not covered by Medicare Parts A and B. This is accomplished through providing health coverage premium plans for beneficiaries enrolled in Medicare Advantage.
D. Prescription drug coverage—insurance provided by private companies to provide Medicare prescription drug coverage to everyone with Medicare. Medicare beneficiaries may purchase prescription drug coverage for outpatient prescriptions.

Medicare is funded by general government revenues and taxpayer funds from such sources as employee payroll taxes, employers, self-employed individuals, and beneficiary plans. Medicaid, on the other hand, is a need-based program funded jointly by federal and state

governments and administered separately by each state government. Medicaid provides health coverage for short- and long-term services for low-income people. Each state defines Medicaid eligibility and administers payments for health care services. Eligible groups include children, families, seniors, and people with developmental and/or physical disabilities. The federal government pays each state according to a formula established by law, which can amount to up to three-fourths of the cost the state pays to provide coverage for Medicaid beneficiaries (Centers for Medicare and Medicaid Services, 2014; U.S. Government Accountability Office, 2011).

The U.S. Government Accountability Office (2011) considers Medicare and Medicaid to be high-risk programs due to their size and complexity, as well as their vulnerability to improper payment (overpayment or underpayment of funds to health care entities) and mismanagement of records. In fiscal year 2010, the U.S. Department of Health and Human Services (HHS) reported that Medicare and Medicaid had about $70 billion in improper payments (U.S. Government Accountability Office, 2011). The Centers for Medicare and Medicaid Services within HHS is leading the effort to reduce the number of improper payments. The Center for Medicare and Medicaid Services is responsible for administering both Medicare and Medicaid and utilizing a variety of technology-based solutions to detect improper payments in an effort to prevent such payments before they are made.

Among the solutions are the Integrated Data Repository (IDR) and Program Integrity (One PI) system. The former is intended to provide Medicare and Medicaid with a single source of data related to their claims whereas the latter is a web-based portal and a collection of analytical software tools used for analysis of data extracted from IDR. Although the Centers for Medicare and Medicaid Services has made many improvements to IDR and One PI in order to achieve its goals, it is not yet capable of identifying and measuring whether these solutions have

provided any financial benefits due to limited use of the system, insufficient data for measurement, and the scattering of data across different state Medicaid programs (U.S. Government Accountability Office, 2011).

Health care fraud has attracted the interest of researchers during the 10 years. The need to address the issue of fraudulent billing transactions by health care providers has led to the exploitation of the modern U.S health care system. This has created a need for data-mining tools and health care system improvements. In 2009, the Centers for Medicare and Medicaid Services estimated health care fraud to be between 3% and 10% of total health care expenditures ($2.6 trillion). The Federal Bureau of Investigation (n.d.) estimates that health care fraud costs American tax payers $80 billion a year. The National Health Care Anti-Fraud Association (n.d.-b) estimates that the financial losses due to health care fraud each year are in the tens of billions of dollars. The incidence of health care fraud continues to rise due to the complexity of the U.S. health care system and the amount of data involved. Failure to implement an effective environment will continue to impact health care costs. Health care fraud is a crime that consists of misrepresentation of facts or providing false information to deceive the health care system for illegal gain (Blue Cross and Blue Shield Association, n.d.; National Health Care Anti-Fraud Association, n.d.-a). This is particularly important in the case of Medicaid because it is operated by state governments individually, not by the federal government as in the case of Medicare.

The federal eligibility requirement for Medicaid states that applicants must be U.S. citizens with low income (Centers for Medicare and Medicaid Services, n.d.). Every state maintains its own Medicaid eligibility guidelines for individuals and families with low income and limited resources. Some states provide eligibility for individuals and families below the poverty line and dictate limited asset resources of a maximum of $2,000. Other states assess

limited asset resources in specific (North Carolina Division of Medical Assistance, n.d.). For example, vehicles are considered a type of asset. A state might allow owning one vehicle regardless of value and evaluate supplementary vehicles to determine eligibility whereas other states might consider the value of one vehicle. Many states do not check out-of-state databases to determine comprehensive asset ownership in their verification process.

Health care fraud detection research focuses on fraud claims by providers, physicians, hospital services, and so on to detect fraudulent billing, abuse, and waste. Data-mining techniques to detect health care fraud have helped detect fraud patterns by providers and reduce the amount of waste and abuse in the health care system. Although researchers (e.g., He, Wang, Graco, & Hawkins, 1997; National Health Care Anti-Fraud Association, 2002; Pflaum & Rivers, 1990) have suggested that most health care fraud is caused by providers, applicants and health care beneficiaries may also account for a large portion of health care fraud. However, less attention has been paid to Medicare and Medicaid fraud by applicants in the health care system. Thus, this research was focused on detecting fraud by applicants in the Medicaid portion of the health care system.

1.2 Problem Statement

The fundamental challenge facing health care today is that it is imperative to detect health care fraud before it occurs while simultaneously providing health care services for those in need. Increasingly, U.S. states are recognizing the need for a robust verification system for Medicaid applicants. Effective, real-time communication between state government resources, federal government resources, and third-party private sources is vital to eliminating health care fraud. For example, some health care services rely on paper records to verify applicants' information. However, this process of verification is unreliable, ineffective, and vulnerable to forgery, loss of

documents, and failure to track changes in applicants' information. Medicaid records have similar vulnerabilities as they rely on paper, which results in unorganized data that cannot be evaluated or analyzed. Medicaid requires a platform that supports the integration, development, and automation of fraud detection of Medicaid applications. The proposed data-driven system combines comprehensive, standards-based Medicaid eligibility guidelines (North Carolina Department of Health Human Services, n.d.-a) with a robust set of fraud detection workflow processes to filter fraudulent Medicaid applications.

CHAPTER 2

Related Work

2.1 Data Mining

Generally, data mining is the process of analyzing a large amount of generated data and summarizing it into useful information that can be used to increase revenue, reduce costs, or both. Technically, data mining is the process of finding correlations, trends, and patterns among several entities in large databases (Milley, 2000). Predictive analysis models derived from data-mining techniques and methods have helped the financial, telecommunications, and health care industries detect fraud. Data-mining models to identify fraud in other domains have also been proposed (Fawcett & Provost, 1997; Ghosh & Reilly, 1994; Grosser, Britos, & García-Martínez, 2005).

Bakar, Mohemad, Ahmad, and Deris (2006) presented the results of an experimental study of outlier detection techniques and indicated that the control chart technique is better than the linear regression technique for outlier data detection. They also described the use of Megaputer Intelligence's PolyAnalyst software for clustering, regression models, and decision trees to fight against fraud schemes. Although these tools do not eliminate fraud before it enters the system, they can help process large volumes of data to detect unusual behaviors. Koh and Tan (2011) and J. Yang (2006) highlighted the limitations of data mining and discussed future directions for research.

Phua, Lee, Smith, and Gayler (2010) conducted a survey to examine fraud detection from a practical, data-oriented, performance-driven perspective rather than the typical application-oriented or technique-oriented view. However, they did not propose a model for eligibility fraud analysis or provide live data set implementation for fraud detection. The Survey was conducted

as a literature review of many fraud detection methods. Viaene, Derrig, and Dedene (2004) combined the advantages of boosting and the flexibility of the probabilistic weight of evidence scoring to explain and effectively diagnose automobile insurance claim fraud. However, the framework applied for diagnosis of automobile insurance claim fraud may not be sufficient for detecting fraudulent activities among applicants for Medicaid benefits.

Shan, Jeacocke, Murray, and Sutinen (2008) applied association rule mining to the examination of billing patterns to detect suspicious claims and potentially fraudulent applications. They identified both positive and negative association rules from specialist billing records. All of the rules were classified as either compliant or noncompliant. Thiruvadi and Patel (2011) discussed effective uses of different data-mining techniques to detect and prevent four different types of fraud: management, customer, network, and computer. Whereas Shan et al. (2008) and Thiruvadi and Patel (2011) concentrated on fraud that already existed in the health care system, this study was focused on preventing fraud from entering the health care system. If fraudulent schemes are targeted before they happen, the amount of federal funds disbursed to fraudsters will be reduced, and the need for the application of fraud detection methods to large data sets will be lessened.

The concept of clinical pathways was initiated in the 1990s for diagnosis and therapeutic intervention by physicians and nurses (Healy et al., 1998; Ireson, 1997). The application of clinical pathways is an efficient approach to analyzing and controlling clinical care in a framework of data mining for fraud detection. W. Yang and Hwang (2006) proposed a framework that was evaluated using a real-world data set to verify the data analysis process on health care fraud and abuse. The experiments showed that the proposed detection model could efficiently identify fraudulent and abusive activities by providers that manual detection models

could not detect through the clinical pathway concept. The detection model serves best as an example for which Bruggemann, Wijma, and Swahnberg (2012) could have used in health care abuse analysis.

Copeland (2011) applied a fraud detection approach using an unsupervised data-mining technique to flag companies with irregular medical claims. By creating and using 12 statistical variables in the data set, Copeland identified six companies that required further investigation for fraudulent activities. The flagged companies all had higher scores than the other companies. The applied fraud detection approach successfully detected 5.9% of flagged companies' net payment in suspicious incontinence supplies claims for Medicaid patients. This unsupervised data-mining approach has been used in the past and will continue to be used for Medical fraud detection because it can efficiently capture patterns hidden in data claims. However, the technique may become less efficient or less reliable as data set volume increases.

Expensive health care costs affect both government health care systems and private health insurers. Allowing health care providers to defraud the federal and private health system only makes it worse. W. Yang and Hwang (2006) focused on a data-mining framework utilizing clinical pathways to facilitate automatic and systematic construction of an adaptable and extensible detection model based on the work of Hwang, Wei, and Yang (2004) and Wei, Hwang, and Yang (2000). Their approach was evaluated using a non-U.S. health care database data set. However, the framework is still relevant because it was tested on a real-world data set. It successfully detected cases of fraud and abuse in Taiwan's National Health Insurance system, though the results were compared to a manually constructed detection model and not an automated detection algorithm model. The detection model outcome demonstrated that the approach should be expanded to handle more noisy data and should be scalable with flexibility to

accommodate health care policy changes. Although the data-mining framework did not uncover all health care fraud situations, it uncovered some fraudulent activity, which was the purpose of creating the model (Chan & Lan, 2001).

Although the health care environment yields a high volume of rich information, it lacks ways of showing hidden relationship trends in its data. Describes each classification data mining techniques; Rules set classifier, IF conditions Then conclusion, Decision tree algorithms, Neural Network Architecture, nuero-fuzzy, and Bayesian Network Structure Discoveries to their application in health care. "If_then_rule" illustrated for the diagnosis of level of alcohol in blood, can be applied in the health care system as a prediction rule to represent a high level abstraction in knowledge discovery according to (Srinivas, Rani, & Govrdhan, 2010). Srinivas et al. (2010) illustrated how each data-mining technique applies to the health care system and presented a method of predicting heart attacks using data mining. By extracting patterns from data warehouses for heart disease to calculate significant weightage patterns, the researchers determined a threshold for predicting heart attacks.

Data-mining challenges are one of the key issues facing health care fraud detection (El-Sappagh, El-Masri, Riad, & Elmogy, 2013). Such challenges include the following (Canlas, 2009; Q. Yang & Wu, 2006):

- Algorithms—very high algorithmic accuracy is needed because health care deals with life-or-death issues. Algorithm accuracy depends on data consistency and can be affected by noisy or missing data.
- Status—data mining must be active with two types of triggers: one to trigger data-mining techniques and one to enforce discovery knowledge within information systems.

- Comparison—data mining must apply techniques then compare the results for selection of the most interesting.
- Results—data-mining system results must be appended to the existing knowledge base.
- Longitudinal, temporal, and spatial support—data-mining techniques must be advanced in order to address electronic health care records (Hripcsak, Knirsch, Zhou, Wilcox, & Melton, 2011).
- Database—data mining must extend beyond a relational database. Although relational databases are the most common type, they must be extended to object-oriented databases and multimedia databases for use with KDD (knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, and business intelligence) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).
- Environment—data-mining distributed environments present a challenge in mining across multidatabase and multirelational data-mining sources.
- Integration—data mining faces the challenge of system integration of visualization tools and database management systems.

2.2 Neural Network

A neural network is a computer program that operates in a manner that is analogous to the natural neural network in the brain. The primary function of neural networks is to emulate the brain's pattern-recognition skills. Li, Huang, Jin, and Shi (2008) provided an overview of all types of fraud in the health care industry and the health care fraud-detection categories applied for the use of statistical methods. Statistical methods are divided into two categories: supervised

and unsupervised. The types of methods used in health care include neural networks (He et al., 1997; Nolting, 2006; Ortega, Figueroa, & Ruz, 2006), decision trees, associate rules, Bayesian networks, and genetic algorithms (Bentley, 2000). Li et al. (2008) describes the data-processing steps: goal setting, data cleaning, handling missing values, data transformation, feature selection, and data auditing, for anlayzing health care data. Algorithms are used in the health care decision tree. W.-S. Yang & Hwang (2006) used the C4.5 algorithm for co-training decision tree method to identify service provider's fraud of the Bureau of National Health Insurance (NHI) in Taiwan.

Travaille, Müller, Thornton, and Hillegersberg (2011) showed that supervised techniques are necessary for an effective fraud detection system. Furthermore, the researchers proved that the techniques in various domains were effective for fraud detection. As a result, no one technique—supervised or unsupervised—can be used to discover all instances of fraud. A fraud detection system consists of multiple techniques to effectively combat fraud and abuse. Becker, Kessler, and McClellan (2005) used a patient sample to rate neural network prediction accuracy relative to binary regression. This technique offers great promise for information prediction.

2.3 Big Data

The term *big data* is used to describe the exponential growth, availability, and use of information, both structured and unstructured. Much has been written on the big data trend and how it can serve as the basis for innovation, differentiation, and growth. Peng et al. (2006) found two types of clustering methods: SAS EM and CLUTO. The researchers used a large health insurance data set to compare the performance of the two methods. Experimental results indicated that CLUTO was faster than SAS EM, though SAS EM provided more useful clusters than CLUTO. The researchers recommended using classification algorithm to predict reliable insurance claims. They also presented the results of an experimental study of outlier detection

techniques and compared two such techniques using a statistical approach with linear regression and control charts. The results indicated that the control chart technique was better than the linear regression technique for outlier data detection. Finally, Peng et al. (2006) analyzed Manhattan distance technique based on distance approach (Bruggemann et al., 2012). Health care industry experts have estimated that if the U.S. health care system uses big data creatively and effectively to drive efficiency and quality, it can generate an annual health care savings of more than $300 billion (Institute for HealthCare Consumerism, n.d.).

One of the essential elements of detecting health care fraud is utilizing an efficient and accurate health care data management system. Any minor problem in data management can make even the most ideal fraud detection models useless. A close look at U.S. health care information systems illustrates four data management problems in the health care system and offers a few insights into future health care system development (Dolins & Kero, 2006; Khosrow-Pour, 2006). Health care data management problems such as data integration issues between heterogeneous systems can affect the outcome of any detection model. For example, if a proposed model relies on inconsistent data integration from federal, state, and private sectors, then it is impossible to associate an applicant between these heterogeneous systems to detect any fraud by the applicant in a state's Medicaid health care database. A recommended solution is the extract, transform, load (ETL) process in data warehousing to accommodate for heterogeneous systems. However, electronic medical records (EMRs) and national data repositories must accompany ETL in order to solve health care data management problem. Utilizing EMR and national data repositories ensures health care fraud detection frameworks receive valuable data input for further analysis and knowledge discovery (Dolins & Kero, 2006; Khosrow-Pour, 2006).

2.4 Data Analysis

Abuse of the health care system is one type of fraud that requires a widespread data analysis. Walker and Avant (2005) developed a concept analysis on health care abuse by utilizing a database index of nursing, Medline, Allied Health literature, and Google Scholar to locate articles on abuse in health care. The result of the concept analysis was that patients' experience with the health care system led to abuse in health care. Patients who felt that their value as a human being had suffered were often unintentionally abusing health care. Thus, health care abuse by patients may be seen as linked to satisfaction with health care coverage and not fraud. The concept of abuse in health care should be taken into consideration by providers and facilities. Bruggemann et al. (2012) proposed a method or technique for investigating the operation of abuse in health care by patients by gathering patient analysis of health care abuse.

2.5 Scoring Model

A scoring model approach is another key technique in detecting medical fraud. Identifying anomalies can provide an effective way of locating hidden fraudulent transactions in health care data. Shin, Park, Lee, and Jhee (2012) proposed a scoring model based on profile information retrieved from electronic health insurance claims to detect abusive billing patterns. The model consisted of two functions: (a) quantifying the degree of abuse and (b) segmentation of providers with similar patterns. The proposed research model was applied to a Korean internal medicine clinic and a national health insurance corporation for outpatient claims. The authors compared the composite degree of anomaly score formulated for intervention and nonintervention groups and examined confusion matrices by intervention history and group to assess the validity of the model. The results showed 38 abusiveness indicators for separate clinics, which were further segmented into homogenous clusters based on their pattern using a

decision tree approach. As a result, the validation of the proposed model was in line with manual detection techniques to identify potential abusers.

The scoring model approach is not limited to just detecting hidden patterns in health care. It has the potential to detect fraud by applicants in environments with many variables and parameters. Agrawal, El-Bathy, and Seay (2012) initiated an integrated data broker services architecture approach to detecting Medicaid fraud at the time of the application approval process. This architecture took advantage of several public databases available through different government and public organizations through a scoring model mechanism and utilized a customizable weighting scheme to determine eligibility for Medicaid services. Research in other fields such as accounting looked at scoring model techniques to predict fraudulent behavior.

Researchers (e.g., Beasley, 1996; Dechow, Sloan, & Sweeney, 1996; Dunn, 2004) have investigated the relationship between corporate government features and financial statement fraud. Financially related warning variable have been investigated by Beneish (1997), Dechow et al. (1996), and Summers and Sweeney (1998). Dechow, Ge, Larson, Sloan, and Investors (2007) developed a model to estimate misstatement probability as a function of accruals quality, market related fraud, and performance measures. The researchers used data issued by the U.S. Securities and Exchange Commission on accounting and auditing enforcement releases to detect accounting fraud through variables identification that correlate with accounting results fraud score output as a screening device to signal further investigation. Dionne, Giuliano, and Picard (2009) developed a scoring model approach to detecting fraud that included insurance fraud detection by using explicitly described fraudulent behavior without limiting the scoring model approach to purely a statistical approach that identified fraud signals and produced fraud probability estimates.

2.6 Layered Architecture

Electronic management of Medicaid applications plays an important role in state departments of health and human services, as it does in the overall U.S. health care system. Health care informatics is being developed to better manage and increase the study of health care. Insight into developing and deploying EMRs and national data repositories to manage health care information systems is offered in Dolins and Kero (2006) and Khosrow-Pour (2006). One insight is the ETL data-warehousing solution to accommodate for heterogeneous systems. Utilizing private asset data resources, such as EMRs and national repositories, is ideal for providing health care fraud detection frameworks with accurate data for analysis and discovery.

Enterprise architecture has also been reviewed as a way to resolve health care data management problems. DePalo and Song (2012) proposed the interoperability of enterprise architecture for health care organizations. By embracing external entities in enterprise architecture for health care interoperability, health care organization can increase patient satisfaction, accumulate meaningful data, and better support business processes. Attention to external entities and data exchange can be used to assess truthful fraud detection tools for fighting against health care fraud.

A hospital case for modeling health care through enterprise architecture provides insight into processing health care–IT integration (Ahsan, Shah, & Kingston, 2010). It incorporates a developed enterprise architecture framework called ArchiMate into a health care reference model to provide an IT service foundation for adapting system design and implementation for health care. (Ahsan et al., 2010) also presented an analysis and overview of health care organization processes in enterprise architecture in their case study. Interrelated components within each layer

play an important role in enterprise architecture since applicants and health care patient concepts may cover many business aspects and application layer components.

Many view web services in enterprise architecture as providing efficiency and optimization in health care. Some use web services for health care fraud detection. The iWebCare platform project is an integrated web service platform for fraud detection for government health care services. The platform design and development provide flexible, online fraud detection modules. The detection of suspicious records across health care system data sets in Europe demonstrates the equality and consistency of the system for fraud detection. The reporting module of the iWebCare platform is responsible for generating and presenting post validation reports in a user-friendly format. Fraud detection is associated with the user interface and informs the user according to behavioral rules once the module discovers suspicious or erroneous records (Tagaris et al., 2009). Although health care literature does not appear to include any published papers demonstrating Medicaid interoperability in layered architecture or a web platform for filtering Medicaid fraud and erroneous Medicaid applications, these concepts are presented here.

2.7 Fraud Detection

Bolton and Hand (2002) reported that types of fraud increase dramatically with the expansion of modern technology. Pattern-detection behaviors are quickly becoming obsolete due to rapid changes in behavior. An ideal proposed model for eliminating fraud in Medicaid must be flexible, scalable, and easy to use in order to eliminate Medicaid fraud at an early stage of Medicaid application, during eligibility determination. The Medicaid Eligibility Application system presents other opportunities for fighting the escalation of Medicaid fraud. For example,

the key issue with building fraud detection tools is adaptation to legitimate and fraudulent behavior changes.

Signature-based predictive tracking can be used for fraud detection in medical transactions. The broad signature-based predictive tracking concept can predict transaction behavior, which is potentially valuable for many applications (Cahill, Lambert, Pinheiro, & Sun, 2002; Cortes & Pregibon, 2001). For example, signature-based tracking can be utilized in Medicaid fraud detection to detect Medicaid applicant behaviors. It is particularly relevant because medical information is stored in homologous systems through penetration of Medicaid eligibility across states.

Li et al. (2008) and Phua et al. (2010) conducted studies of fraud-detection techniques from a practical, data-oriented approach to detect fraud in health care, electronic fraud, and fraud in other industry areas. Data-mining techniques of statistical methods applied to health care for fraud detection include decision trees, neural networks, association rules, Bayesian networks, and genetic algorithms. These techniques have recovered millions of dollars of U.S. health care funding and captured many fraudulent providers, facilities, and organized entities. Furthermore, fraud scams regenerating from "hospital stay conflict, hospital stay with no associated physician inpatient visit, excessive lab/radiology services per client per day, X-ray duplicate billing, fragmented lab and X-ray procedures, lab/X-ray interpretation with no associated technical portion, and ambulance trips with no associated medical service" can now be detected (Li et al., 2008; Sokol, Garcia, West, Rodriguez, & Johnson, 2001).

Data-mining fraud-detection techniques must be used to analyze data from the health care system, a complex structure to detect unknown patterns in the data. SAS (King & Malida, n.d.), Exodus Payment System (Exodus Payment Systems, n.d.), and Dun and Bradstreet (Mears &

Dun & Bradstreet, 2012) provide alternative methods for detecting Medicaid fraud, such as the eligibility fraud method, the broad data source pool approach, or the use of biometric engine technology to tackle a small area of the big problem. Utilizing different tools from various companies to fight against each type of fraud creates inconsistency, inefficiency, and unreliability in terms of fraud detection in the health care database system. However, utilizing one tool to detect all fraud types creates an ideal solution to the health fraud problem. Health care industry could benefit from a software solution that integrates fraud detection techniques for outlier detection with data mining, clustering, and predictive models to solve the Medicaid fraud problem in all four categories: processing, organization, technology, and analytics.

Although there are many software packages, tools, and methods for detecting fraud and recovering millions of health care federal dollars, they come at a cost. These costs continue to increase as fraud-detection data-mining software and analytics tools require more integration in order to detect new fraud. For example, unintegrated software is ultimately more expensive than a fully integrated software solution.

Two case studies provide an example of utilizing a fully integrated software solution to detect fraud: the New York State Department of Taxation and Finance case study on income tax refund fraud, abuse, and debt collection and the North Carolina Department of Health and Human Services case study on Medicaid fraud detection. The first study consisted of 5,000 employees, about $60 billion in annual income collected, and taxpayers from a wide range of demographics and cultural backgrounds. The New York State Department of Taxation and Finance applied IBM's Integrated Business Solution and predictive models to identify the next-best audit selection. As a result, the department's revenue increased by $889 million in the first

five years. The system also “increased screener and auditor productivity,” “enhanced taxpayer correspondence,” and “improved audit program management.” The second case study consisted of 100 Program Integrity Unit employees, about $14 billion in annual paid claims, and $25 million in recoupment letters issued each year. The North Carolina Department of Health and Human Services incorporated IBM’s health analytics solution to detect suspicious patterns in claims and identify suspicious providers in real time as they filed claims. As a result, the North Carolina Department of Health and Human Services recovered between $60 million and $100 million over a 12-month period, identified $140 million in claim data, recovered $86 million of the $555 million in personal care services, and recovered $55 million of the $235 million in durable medical equipment.

IBM’s sophisticated and real-time analytics software provides proper analytical oversight. Using complex mathematics and model statistics to examine existing data sources in a faster, smarter, and better way allows states to achieve a positive return on investment, greater efficiency, and greater fraud detection reliability (IBM Corporation, n.d.; North Carolina Department of Health and Human Services, n.d.-b).

Many fraudulent techniques emerge every year to illegally gain health care funds and benefits. Furthermore, as data volume increases and Medicare and Medicaid expand, it becomes necessary to investigate this problem from every aspect and dimension. IBM’s Smarter Signature Solution provides the following solutions to the health care problem;

1. detecting suspicious transactions before payment is arranged
2. minimizing loss from fraud over payment
3. analyzing a range of suspect behavior from claims and providers simultaneously
4. analyzing and flagging past suspicious patterns

5. analyzing new fraudulent schemes from similar detect fraud activities

6. detecting fraud in real time to stop illegal use of health care funds

These are some of the many services IBM provides. There are many more in IBM's Integrated Software Solution to combat against state health care fraud (Yueh & Barry, 2010).

CHAPTER 3

Methodology

3.1 Current Approach Used By State Health Care Departments

Every state maintains its own Medicaid eligibility guidelines for individuals and families with low income and limited resources. Table 1 provides a sample of state Medicaid eligibility requirements. The requirements by states shown in Table 1 are used to verify applicants' information against federal and in-state database resources. States do not check out-of-state databases or third-party database resources for broader information accuracy. The process starts with checking major federal government databases such as the Beneficiary and Earnings Data Exchange database (Bendex) or the social security database for determining citizenship, matching social security numbers, and checking in-state databases such as the Department of Motor Vehicles (DMV) database to check for any vehicles that the applicants own. Also, states can use bank account documents to determine applicants' asset resources and other assets that applicants declare in their application.

Table 1

State-by-State Eligibility Requirements

Sample State-by-State (Eligibility Requirements for Individuals)				
State	**Poverty level**	**Income**	**Resources**	**Asset specified**
NC	200%	$22,980	$3,000 cash	One vehicle regardless of value
GA	133%	$15,282	$2,000 cash	Vehicle value up to $4500
TN	185%	$20,000	$2,000 resources	Vehicle value
NJ	200%	$22,980	$2,000 resources	Vehicle value
NY	200%	$22,980	$2,000 resources	Vehicle value

3.2 Problems With The Current Approach

Escalation of fraud usually occurs when a wide range of ineffective cooperated health care environments are open for fraud. Examples of Medicaid eligibility fraud include resource

misrepresentation, eligible members sharing resources and benefits with ineligible members, misrepresentation of medical conditions, failure to report third-party resources, and eligibility determination issues. Fraud in general usually escalates when environments have one or more broken areas of detection. Health care fraud in the United States occurs due to health care complexity system and because the health care system relies on other factors such as federal and state systems. Problems in Medicaid fraud should be addressed while factoring in other collaboration systems such as those mentioned in chapter 2.

An opportunity for fraud exists when state agencies do not check out-of-states databases to determine the asset resources of an individual and compare this to what the applicant has filed in the application process. For example, an individual or family may relocate from one state to another and apply for Medicaid benefits in the new state while still owning assets such as vehicles, boats, bikes, and real estate property in the old state. When states only check applicants' resources against in-state databases, applicants who are ineligible because of undeclared out-of- state assets may become eligible for Medicaid benefits.

Health care departments may use assets, bank account statements, and income documents from employers provided by the applicant to initiate the eligibility process. When the chain of documents is broken, applicants may find ways to submit fraudulent documents. For example, an applicant might alter income documents from an employer or bank account statements before submitting them.

Cooperation at the system level presents another key problem in this process. Lack of unification of data sources between states, local state governments, and the federal government creates opportunities for fraud. For example, state Medicaid services departments can gain some direct access to federal database resources for their infrastructure operations. However, it appears

that states cannot gain direct access to other states' resources such as Medicaid services department files, DMV databases, and so on. Slow processes due to manual detection and reverification present other opportunities for fraud. Such manual operations lack accurate results, allow for inaccurate data and information storage, and permit more fraudulent schemes. If the verification of thousands of applicants depends on a human representative, there is an opportunity for white-collar fraud or human error. By contrast, an automated process can yield accurate results and accumulate detailed reports.

Privacy laws may present an opportunity for fraud by applicants and users that health care department representatives must be aware of. Although it is not possible to conceal information retrieval and operational processes from the public, utilizing private and public databases reduces the opportunity for fraud.

3.3 Medicaid Eligibility Application System Prototype

The problems mentioned in the previous section must be addressed by researchers. Although many researchers have used a variety of data-mining techniques to detect fraud, they have focused on providers, false services, and improper billing. The prototype system proposed in this study represents the integration of an eligibility determination system with in-state and out-of-state public asset databases. It involves checking Bendex records for identifying citizenship status, legal residence status, social security records, banks records through tax returns, and public and DMV databases for checking assets. Public records from Data Broker databases (Agrawal et al., 2012) are also used to identify and retrieve assets that an applicant may or may not have. The system is focused on applicants at the beginning stages of Medicaid eligibility application. The following subsections represent the proposed architecture of the system including algorithms and the eligibility process for detecting and eliminating fraud.

3.3.1 Model overview

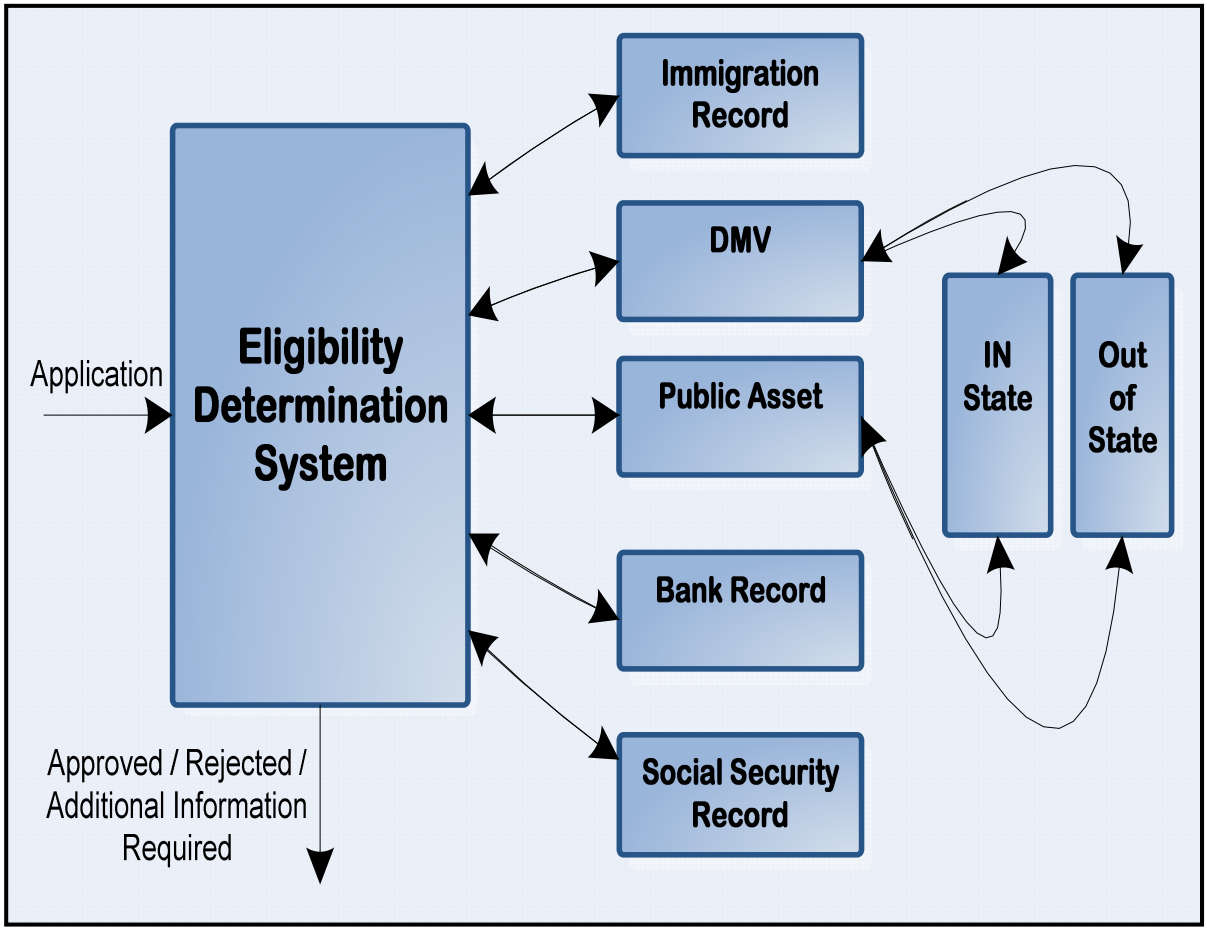

Figure 1 An Integrated Medicaid Fraud-Detection Model

Figure 1 shows the proposed Medicaid eligibility system model for identifying public assets with in-state and out-of-state records. According to the model, an applicant's application information is matched against public records. Each type of category consists of a predetermined weight. Depending on the results returned from these public records and using these weights, a final eligibility score is calculated based on a threshold value. As a result, an applicant's application may be rejected or accepted or additional information may be sought.

3.3.2 Eligibility flow

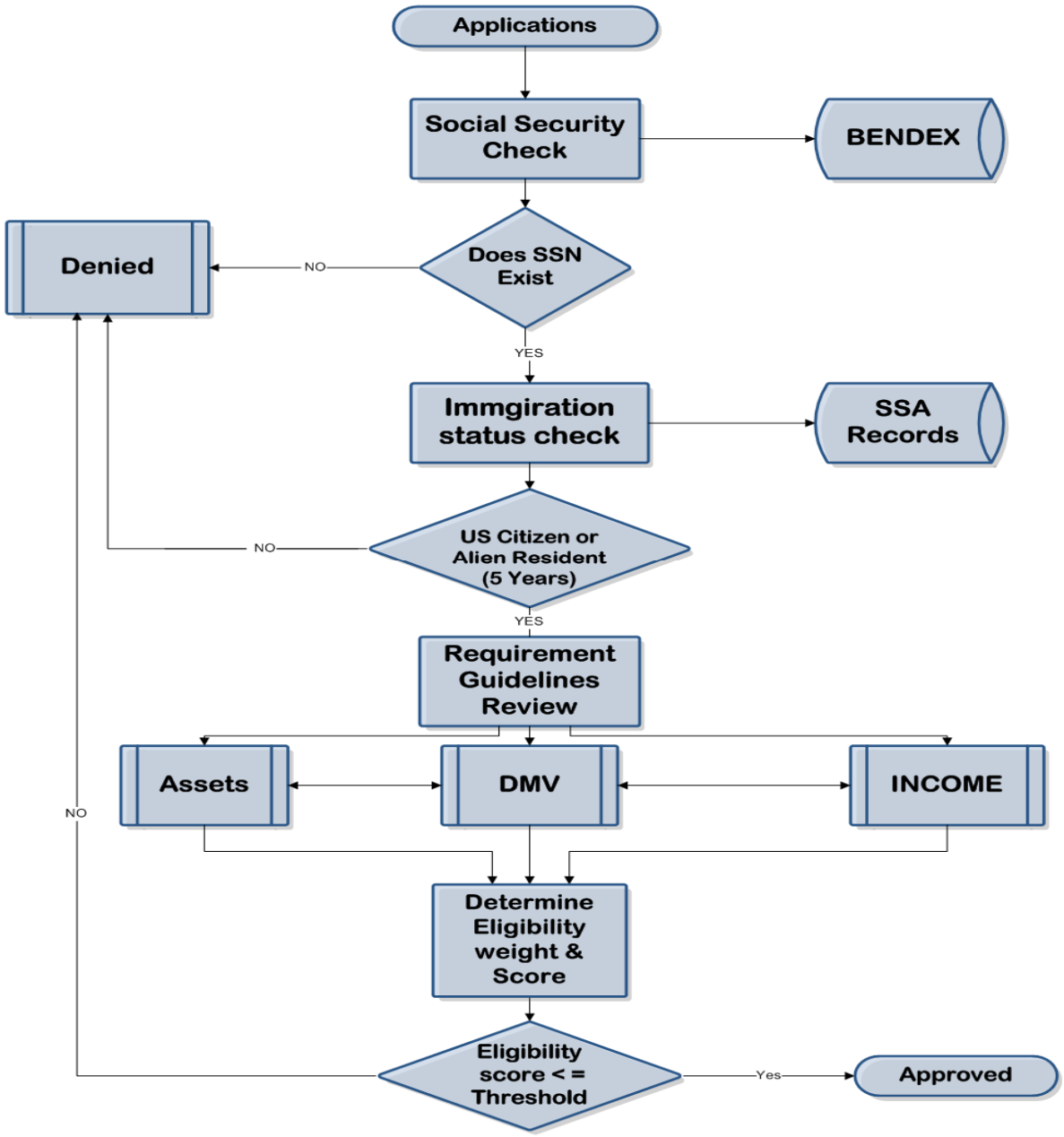

Figure 2 Medicaid Eligibility Flow Model

The eligibility process is initiated when an applicant files an application for Medicaid benefits, providing the system with information such as legal name, social security number, address, marital status, and so on. Figure 2 illustrates the entire process flow from the start of the

application to the decision output. The eligibility determination system sends the data to a process called social security check to compare the applicant's social security number to the Bendex database for match based on the applicant's name and Citizenship verification through Social Security Administration (SSA) records (Social Security Administration, n.d.).

When these two major factors (social security number and immigration status) are verified, the requirement guidelines review process begins. This process comprises a variety of subprocesses to check for eligibility. Each subprocess identifies all assets associated with the applicant from DMV records, income records, and other public asset database resources. The goal is to create an algorithm based on the weights and scores of all assets identified from public records while allowing for scalability based on state requirements. Cost of living, region, and property values differ between states; therefore a unified algorithm that supports all state parameters would be difficult to create, inefficient, and unscalable. Allowing states to adjust the algorithm to their parameters is ideal for detecting Medicaid fraud.

3.3.3 Eligibility determination and calculation. There are many assets that an individual or family may own. Assets are grouped into five categories as follows: (a) equity (e.g., commercial real estate, private real estate), (b) fixed assets (e.g., vehicles, boats, planes, bikes), (c) collectables (e.g., art, coins, stamps, wine), (d) cash sources (cash on hand, cash at the bank in checking or savings), and (e) others (e.g., bonds, stocks, fixed interests). These categories provide the means to retrieve the appropriate score. The categories can include exempt or nonexempt assets. For example, states may exempt one vehicle asset or one vehicle value for applicants who utilize this asset for transportation to doctor appointments, shopping, and personal use. Meanwhile pleasure assets such as boats and luxury vehicles, especially for those who own more than one vehicle are nonexempt.

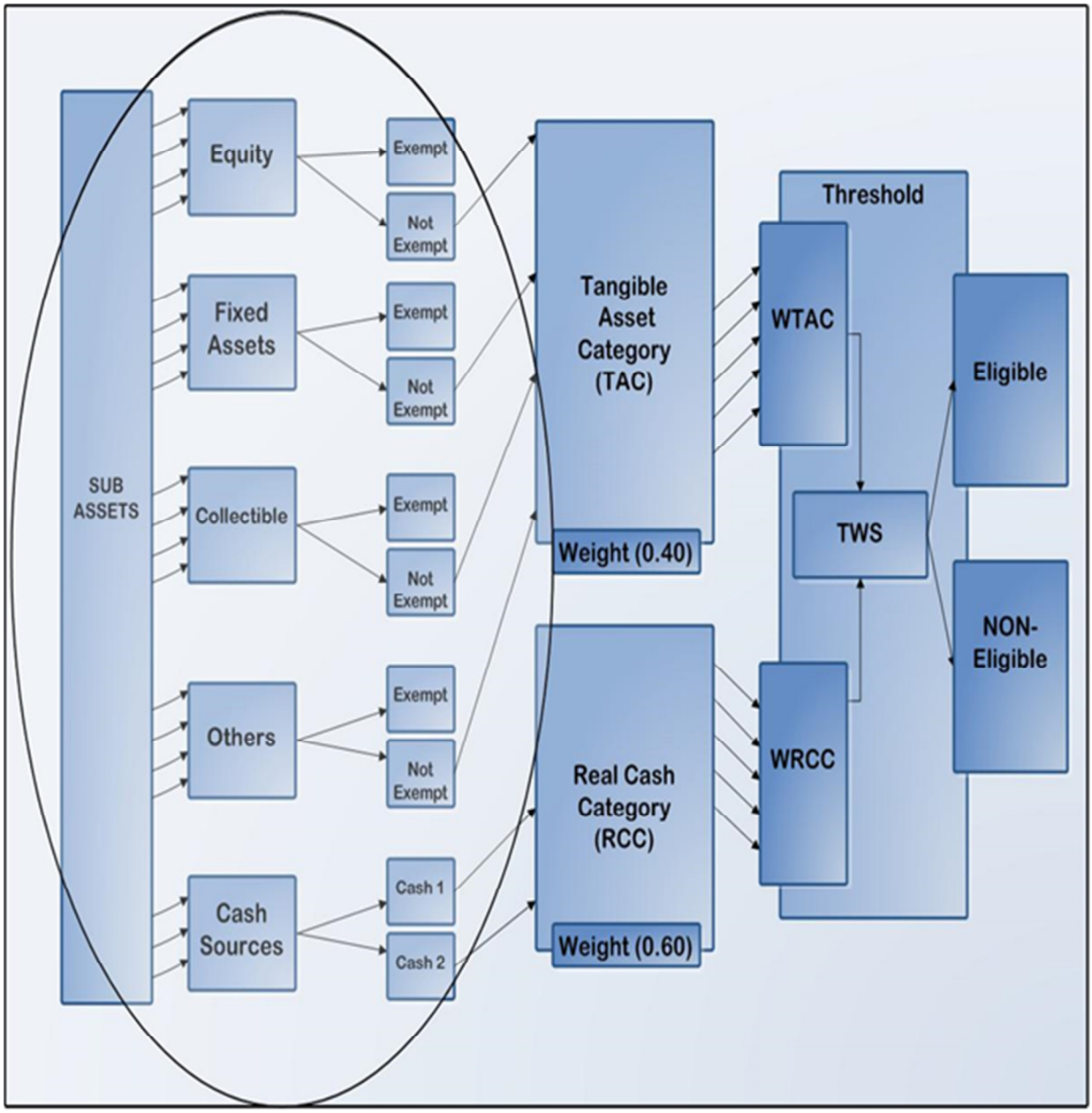

Figure 3 Proposed Eligibility Determination Using Various Asset Categories

Figure 3 shows the identification of all subassets identified, which are then placed into two categories—tangible assets (TAC) and real cash (RCC). Each is assigned a weighted score that will be combined for a total weighted score result (Wscore). This total score is compared to a threshold to determine eligibility or noneligibility. Classifying the subcategories into two

groups (TAC or RCC) allows the use of scoring, weights and the threshold approach in the algorithm.

TAC comprises 10 types of assets in the subasset category whereas RCC comprises five types of assets in the subasset category. Dividing 10 by 5 parameters results in a 2-TAC parameter table. The first parameter has a score of 2, and each subsequent parameter increases by 2. For RCC, dividing 5 by 5 parameters results in 1 RCC parameter. The first parameter has a score of 1, and each subsequent parameter increases by 1. Tangible assets are considered assets that are physical and can be converted into cash within a year whereas real cash includes types of assets that present cash in real time. For example, real cash includes the following:

1. checking accounts
2. saving accounts
3. cash in hand
4. IRA accounts
5. other cash sources such as 401ks, saving bonds, and home equity

Tangible assets include the following:

1. stocks
2. bonds
3. treasury bills
4. investment property
5. vacation homes
6. livestock
7. collectables such as precious metals and coins
8. homes

9. fixed such as vehicles, boats, aircraft, and watercraft
10. IRAs

Each state government can organize and adjust these assets as required by state regulations. All asset types are introduced in the subasset categories, and state governments determine how each type will be classified—either as tangible, real cash, or a third classification if needed. The parameter score is calculated based on how many subassets are included in each category, and the number of parameters required. Tables 2 and 3 illustrate the score for each parameter within the two main categories.

Table 2

Real Cash Parameters

Range	Range Value	Score
A	$0 - $499	1
B	$500 - $999	2
C	$1,000 - $1,499	3
D	$1,500 - $1,999	4
E	>= $2,000	5

Table 3

Tangible Asset Parameters

Range	Range Value	Score
A	$0 - $149,999	2
B	$150,000 - $299,000	4
C	$300,000 - $449,999	6
D	$450,000 - $599,999	8
E	>= $600,000	10

3.3.3.1 Algorithm to calculate weighted score. Classifying assets into two categories (TAC or RCC) yields a weight of 0.40 for TAC and a weight of 0.60 RCC based on the subcategories included. In real implementation, cost of living and state policies determine the

appropriate category weight and range parameters for TAC and RCC scores. Figure 4 represents the integrated algorithm.

Input: Ast_i, Ast_c , TAC_w, RCC_w, ETH

(Ast_i = Individual Asset, Ast_c = Asset Category, Tac_w = Tangible Asset Weight, RCC_w = Real Cash Weight, ETH = Eligibility Threshold)

Output: TA, RC

FOR each asset Asti to be categorized to category parameter

IF (Asti= Equity), or (Asti = FixedAssets), or (Asti = Collectable), or (Asti = Others)

THEN

If Ast_i is of type "Tangible"

TAC = TAC +Ast_i

ELSE

RCC = RCC + Ast_i

END IF

END FOR

TA = TAC * TACw

RC = RCC * RCCw

THEN

TWS = TA +RC

Eligible = TWS <= ETH

Noneligible = TWS > ETH

END

Figure 4 Algorithm 1

The algorithm presented in Figure 4 shows the proposed data-driven approach with accurate means to classify resources into the appropriate categories using Equation 1.

$$\sum_{i=1}^{n} P_i \; w_i \tag{1}$$

Because states differ in terms of their cost of living and policy statues, Algorithm 1 and Equation 1 can be customized to state requirements. For example, a state may elect to factor other weight measures into the resource categories—for instance, by weighting the current gross domestic product for a specific resource acquired by applicants. This allows scalability for future outcomes.

3.3.3.2 Total weighted score. The five categories (entities, fixed assets, collectibles, cash sources, and others) provide parameters for the algorithm based on classification and level of importance. The total weighted score produces two weights from five parameters.

- N = {1,…., n}, the number of criteria.
- i = {1,…., i}, the number of values to be assigned
- P = parameter value for *i*th value
- W = assigned weight for *i*th value

The total weighted score is given by Equation 2:

$$\mathbf{P1 * W1 + P2 * W2} \tag{2}$$

The methodology section described the proposed architecture including algorithms, calculations, and the process for detecting and eliminating Medicaid eligibility fraud. The following chapter describes the implementation of a prototype for such a system, which can be adopted by any U.S. state.

CHAPTER 4

Proposed Data-Driven Implementation

4.1 Layered Architecture

Many states utilize a mainframe for their board of education, health care services, or IT departments. For example, the State of North Carolina utilizes a mainframe for its Department of Transportation and Department of Health and Human Services. However, the evolution of IT to support business needs and/or detect fraud sometimes creates a heterogeneous environment across heterogeneous platforms, which creates a challenging environment for detecting fraud.

The objective is to support mainframe and distributed environments while integrating various platform products to better test the algorithm for effective results. The prototype was designed based on a three-layered architecture. Layered architecture provides many advantages such as flexibility, maintainability, and scalability. Also, by separation of the user interface, business logic, and data access layer, integration concerns can be addressed regarding logical layers and components across federal and state departments. The motivation behind data-driven implementation is to incorporate modern IT into Medicaid business processes. Information associated with each Medicaid applicant should be properly analyzed, queried, stored, and accessed. The ability to retrieve public information is essential to addressing the issue of health care fraud.

The prototype consolidates automated electronic data management of Medicaid applications and sets robust fraud detection workflow processes into one integrated, data-driven infrastructure. The architecture is developed for any state Medicaid department. The implementation is based on three-tier architecture with a web-based Medicaid application platform, which allows for a clear separation between applicants, verification workflow

processes, and data storage. Furthermore, this infrastructure provides state governments the ability to add or replace layers to interoperate different Medicaid services in their health care departments.

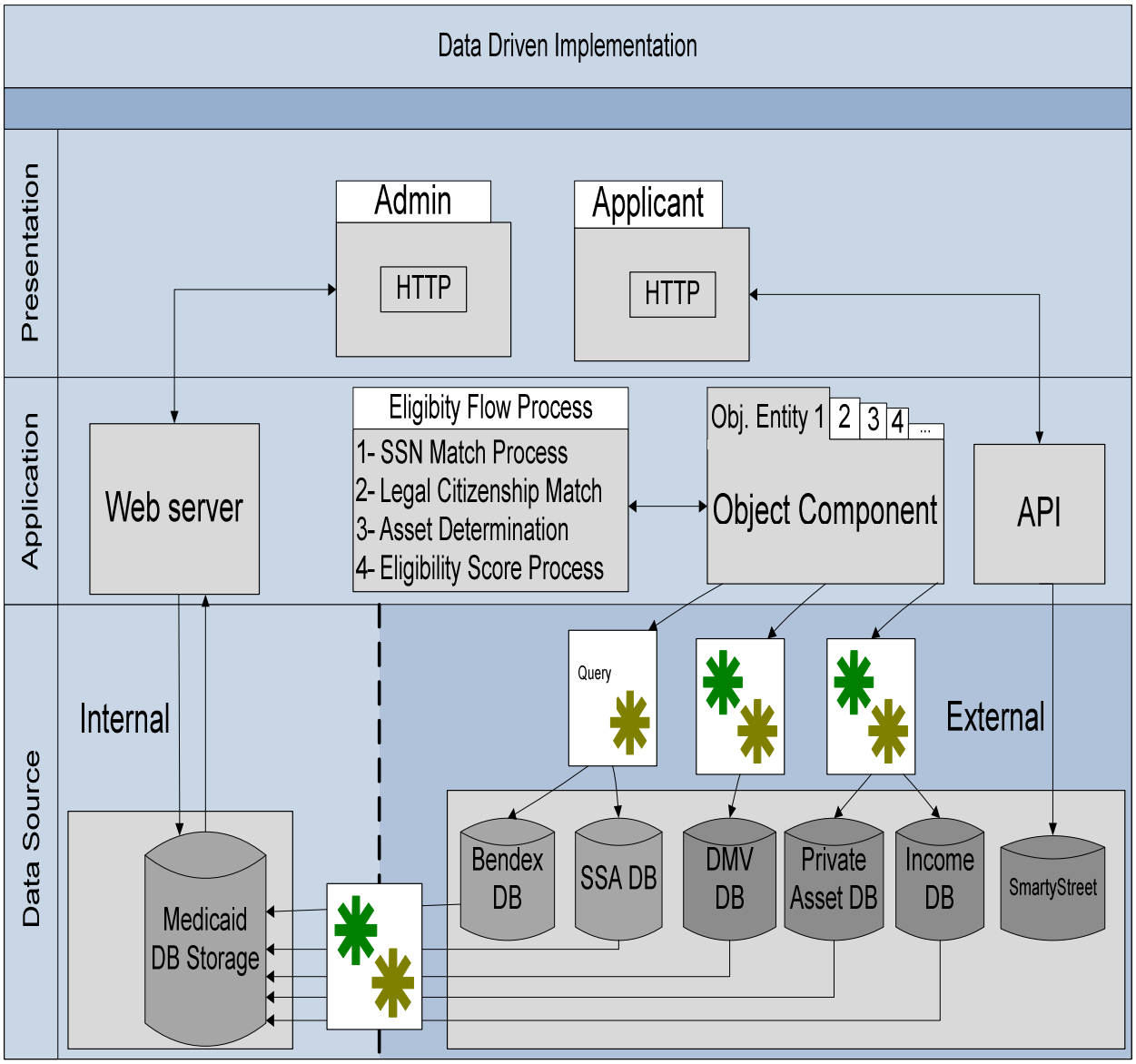

Figure 5 Data-Driven Layer Architecture

Figure 5 provides a logical view of the data-driven layer architecture proposed for implementation. The details of the architecture are divided into the following subsections:

1. Presentation layer—the visible part through which users interact with the system.

2. Application logic layer—the set of robust fraud detection processes.
3. Data source layer—the data association and involvement with the entire infrastructure.

4.1.1 Presentation layer. The presentation layer, also called the graphical user interface (GUI) layer, is the visible section of the architecture. Users interact with the system using this layer through hypertext transfer protocol (HTTP). Typical users are Medicaid applicants who are pursuing Medicaid services, Medicaid representatives who assist in application processing and helping Medicaid applicants who are computer literate or lack access to the HTTP client, and Medicaid Administrators who manage the overall application process.

The presentation layer contains applicants' and health care administrators' GUIs and user forms for the data feed. These forms include the following: (a) sign up, (b) sign in, (c) welcome, and (d) Medicaid application. Applicants see the sign-up form before they can sign in to access the Medicaid application web form. Figure 6 illustrates a sample sign-up form. It includes basic information for associating an applicant with his or her account.

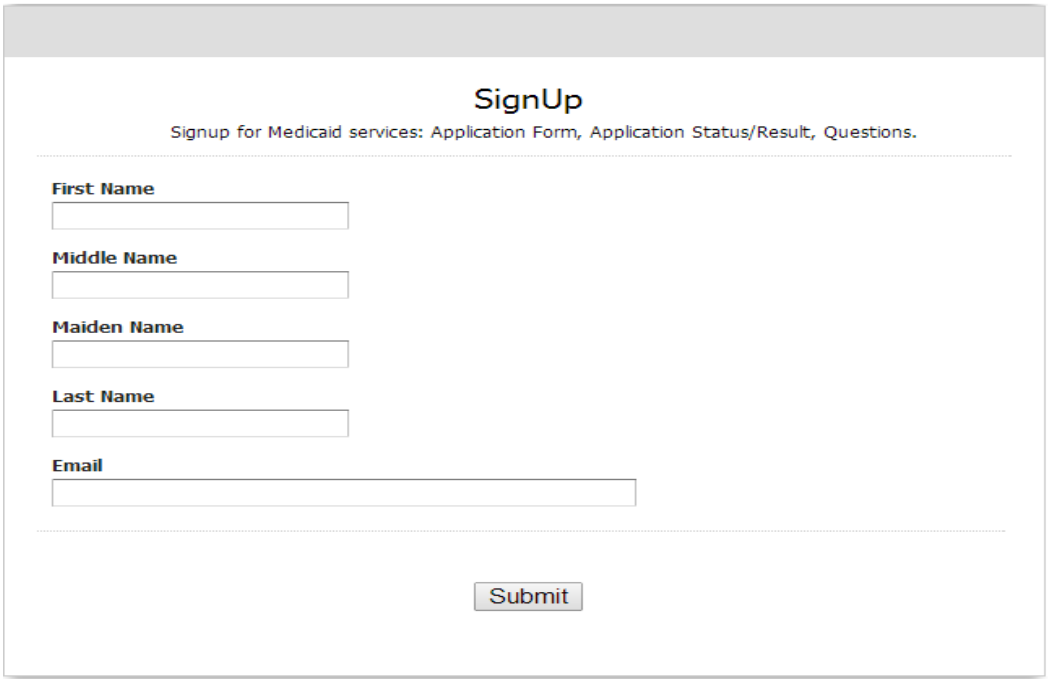

Figure 6 Sign-up Form

By contrast, representatives and administrators log in using predefined credentials to access new and existing Medicaid applications. Figure 7 illustrates the sign-in requirement for accessing Medicaid accounts.

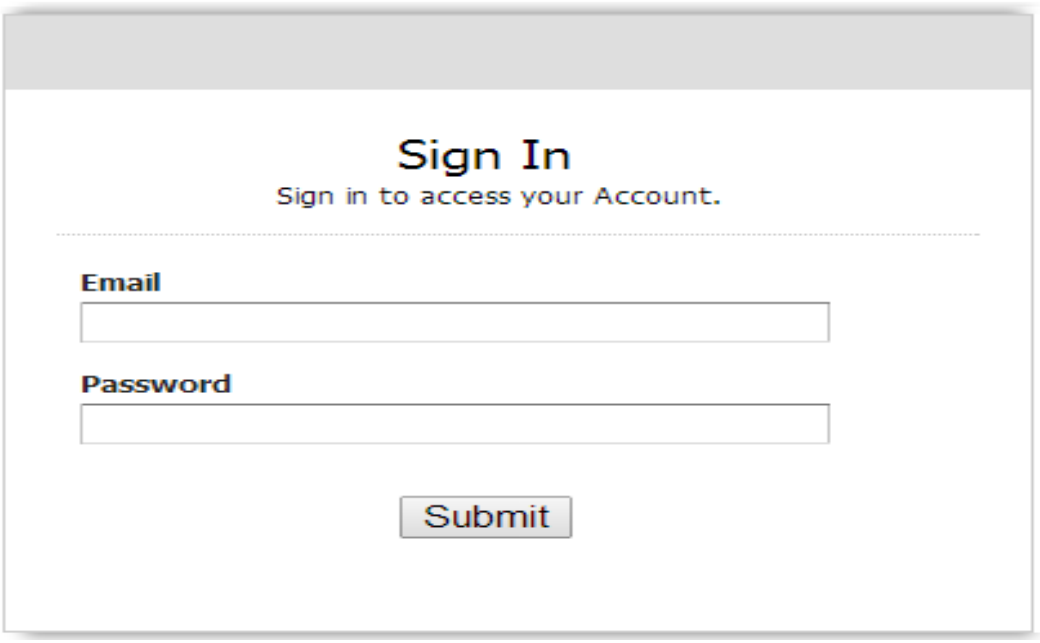

Figure 7 Sign-In Form

Applicants will also utilize this form to access their account. Unlike Medicaid administrators, applicants can only view their own account. Medicaid administrators can view all accounts under review.

The presentation layer includes web browser processes for displaying HTML requests and processing HTML responses. Users can access the site through a variety of web browsers, such as Firefox, Internet Explorer, or Safari. The web browser communicates with a web server using a standard protocol for properly displaying HTML pages on the user HTTP client without the need for prior configuration.

After sign in, the web server transfers and displays the appropriate screen based on the user's sign-in credentials. For applicants, the welcome screen illustrated in Figure 8 displays

three menus for the user: account profile, new application, and check application status. This feature will allow Medicaid applicants to electronically submit their Medicaid application, review application status, and access any messages or notifications about issues with their application via their account profile.

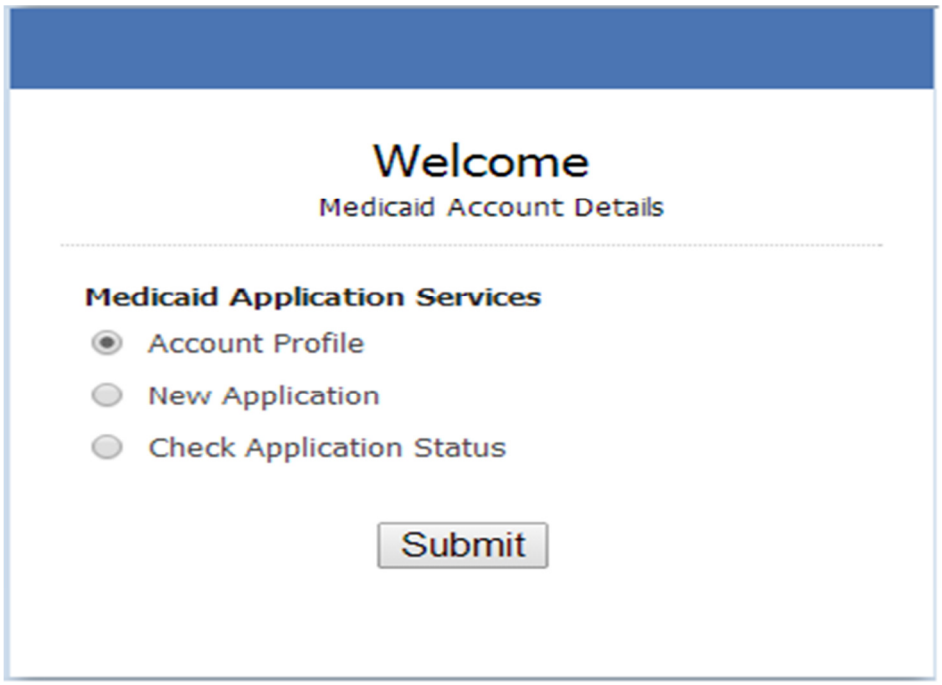

Figure 8 Welcome Screen User Interface

Once the user chooses to the new application menu, the Medicaid application user form is displayed to the user. Figure 9 represents the data-driven Medicaid application web form. It consists of required input fields. The web form is a sample form based on the North Carolina Medicaid application (North Carolina Department of Health Human Services, n.d.-a).

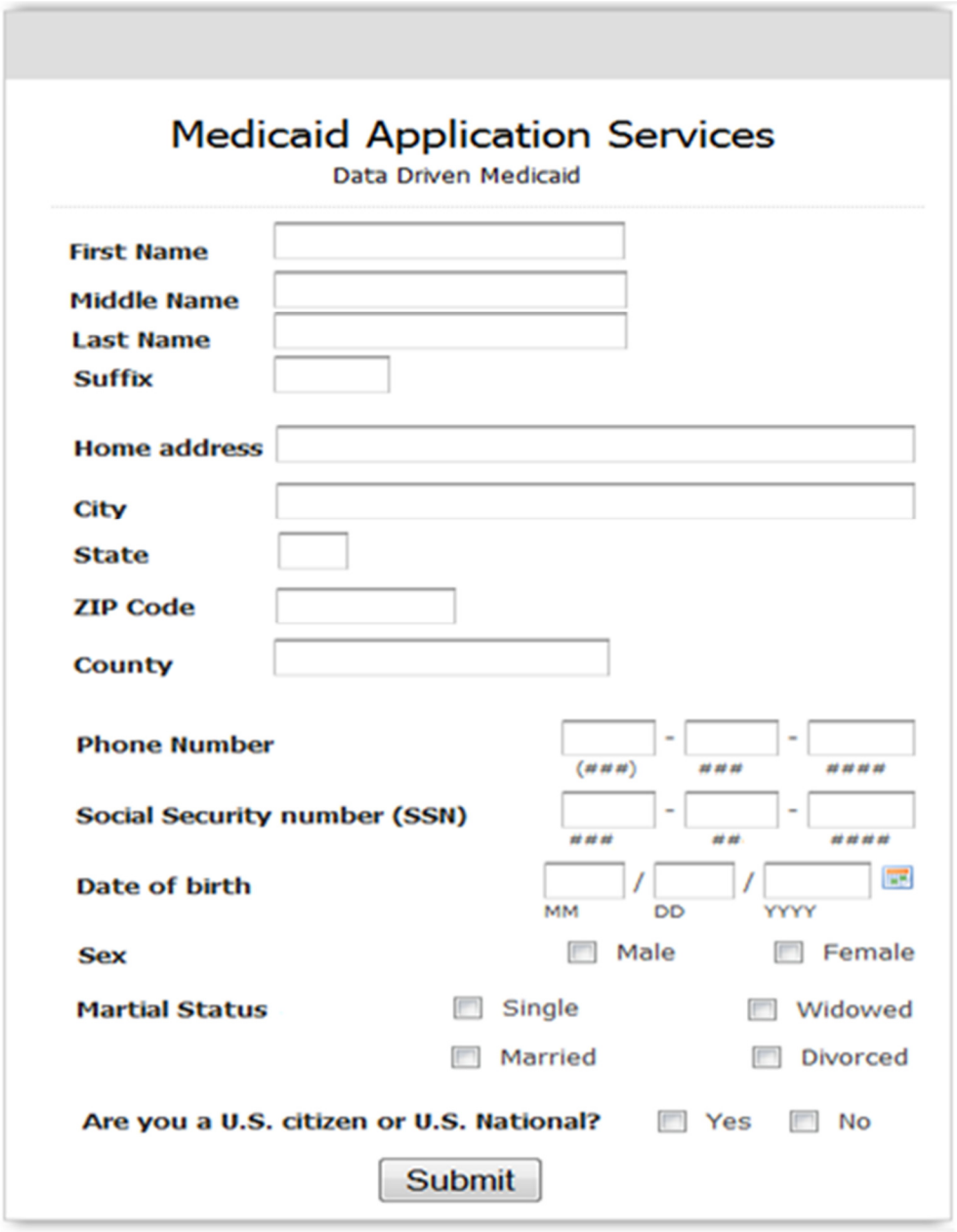

Figure 9 Data-Driven Medicaid Application Web Form

Many of the input fields are streamlined for accurately collecting appropriate data pertaining to the input field. Applicants' home address, city, state, and zip code are verified instantly through the SmartyStreets application program interface (API) for precise information (SmartyStreets, n.d.). The county input field is accordingly populated with the counties of that state. The rest of the input fields are also simplified for data consistency across all applicants.

The main purpose of this layer is twofold: (a) to electronically collect user information, pass it to the application layer for processing, and then reveal a response output such as the application result, whether the applicant is being monitored, and the application status, and to (b) collect precise and consistent data from applicants for future analysis.

4.1.2 Application layer. The application layer is the key structure in enterprise architecture. It acts as the principle for organizing logic flow between business processes and IT data infrastructure, reflecting the integration and standardization requirements of the system model. It is responsible for retrieving, processing, and transforming data. The application layer in this data-driven design consists of an eligibility flow process as the application logic workflow. The application layer manages four processes, the object component for each of the four processes to carry out its instructions, and the object entity components for allocation and distribution of data. As for the application logic workflow, it is responsible for systematically executing a sequence of processes to attain a business process. Once the data are collected from the presentation layer and passed to this layer, the eligibility flow process initiates (a) a legal citizenship match, (b) an income match, (c) asset determination, and (d) the eligibility score process.

The eligibility flow process is concerned with fulfilling each business process in order to ensure that business rules and Medicaid application fraud detection are successful. It is also responsible for processing interruptions in the event of data verification failures or fraud detection. A review of the federal eligibility requirements (Centers for Medicare and Medicaid Services, n.d.) is performed in the first two processes. Asset determination and eligibility verification are performed in the last two processes. This allows the eligibility flow process to interrupt the workflow in the event of fraud detection in an applicant's social security number or

legal citizenship claim to skip the asset determination step to the eligibility score process. Each process in the eligibility workflow calls out a specific object component for that particular process. Object components are the engines behind the process. They include instructions according to business rules for processing. Furthermore, each object component is further divided into an object entity component. Object entity components will capture the necessary data according to the object component. They also store data to the Medicaid database (Medicaid DB Storage) and ensure data consistency according to Medicaid business rules.

Accordingly, after the data are passed to the application layer, the eligibility flow process will initiate a citizenship match process. This is a combination of a social security number match and legal status match. This process will call out the social security number match process component for execution and processing of social security number match instructions. The object entity component within the social security number match process component will evaluate and transform data to store it in Medicaid DB Storage.

Upon successful completion of the social security number match process, the legal status process matches for citizenship. Then, the eligibility workflow moves to the next process: the income match process. The income object entity component retrieves applicants' income and identifies whether it is below the poverty line or not. If the applicant does not satisfy the income and citizenship requirements, the process stops and jumps to the eligibility score process. Figure 3 represents a logical view flow chart of the eligibility flow process, which illustrates the entire process flow from the start of the Medicaid application to the decision output.

Subsequently, the eligibility workflow transfers the process to the asset determination process then the eligibility score process. Then the eligibility algorithm, including total weighted score and weight score calculation, is implemented in the asset determination process. Then we

incorporate the eligibility determination score in the object entity component of the eligibility score process.

4.1.3 Data source layer. This layer provides access to different database types. There are two general database structural models in the health care industry that are used during the application process: hierarchical and relational. The hierarchical database is used in the mainframe's management information system and stores data in inverted format. The structure in the Relational Database Management System (RDMS) stores data, such as binary large objects, XML, and other object-oriented data in rows and tables. A relational database structure was used for implementation testing on Medicaid's internal database and external resource databases.

The internal data source was used to connect to the internal database (Medicaid DB Storage), which is used for storing data from Medicaid applications and application layer processes. These also use an internal database for data retrieval during application layer process implementation. The internal database comprises applicant and spouse information and applicant match schema. Each table schema is discussed in the next section. The external part of the data source represents databases located in different physical tiers for retrieving matching information regarding the applicant. It encompasses the following external databases:

- *Bendex:* The Bendex database is an SSA database for exchanging social security numbers with states agencies on a daily basis through Bendex Connection (Social Security Administration, n.d.).
- *State Verification Exchange System (SVES):* This is another SSA database that includes SVES I, SVES I/Citizenship, SVES II, SVES II, and SVES IV for federal agencies to extract citizenship data of an individual for citizenship verification via SVES service connection (Social Security Administration, n.d.).

- *DMV:* This database corresponds to a local state DMV database. It is used for extracting owners' vehicle information, such as how many vehicle assets the owner owns and the asset's market value. Connection to this database is via a direct link (North Carolina Department of Transportation, n.d.).
- *Private Asset*: A database by KnowX (http://knowx.com) for retrieving aircraft assets, real estate, U.S. Coast Guard vessels, watercraft, and other assets. Connection to this database and data format is uniquely personalized to the type of service required by the user (KnowX, n.d.; LexisNexis, n.d.).
- *Governmental Liaison Data Exchange Program (GLDEP):* This database is used for data sharing between the Internal Revenue Service (IRS) and state tax agencies. It includes taxpayer income information. Connection to this database is via a GLDEP service link (IRS, n.d.).
- *SmartyStreets:* This database provides fast and easy U.S address verification. It validates applicants' addresses by verifying them in real time with the SmartyStreets API. If the address is ambiguous, SmartyStreets displays multiple matches. This allows the applicant to be alerted via HTML when the address is invalid. Once the data are verified or corrected, the JSON file provided by SmartyStreets is extracted and the user address information is updated with the correct syntax. This will provide address consistency in Medicaid DB Storage (SmartyStreets, n.d.).

Bendex is used for matching applicants' social security numbers. Then SVES is used to retrieve applicants' citizenship status. Next, the GLDEP, Private Asset, and DMV databases are used to orderly retrieve applicants' income and assets. The DMV database is used as a secondary

asset verification database. Medicaid DB Storage is used upon retrieval of data from the mentioned databases for data storage and application layer processing.

Because Medicaid data or other health care data sets could not be used for this study due to privacy concerns, a sample of synthetic data was created to represent a live database system. This prototype was used to conduct testing and validate the data-driven implementation. The following database tables are described in detail.

- *Applicant*: This table included a sample of required information about each applicant applying for Medicaid benefits and services. It included personal information about the individual, such as first name, middle name, maiden name, social security number, sex (male or female), date of birth, race (Asian, Black or African American, White or Caucasian, American Indian or Alaska Native, Native Hawaiian or Hispanic Cuban, or Other Pacific Islander), legal status (citizen, noncitizen, or permanent resident), alien registration number, and marital status (single, married, divorced, or widowed).
- *Spouse info*: This table included a sample of the necessary information about the applicant's spouse. It was linked with the applicant schema via ID number. Figure 5 displays spouse's social security number, spouse's first name, spouse's middle name, and spouse's last name. However, the prototype included more attributes of spousal information, similar to the attributes included in the applicant schema.
- *Applicant match*: This table consisted of attributes required by the application layer for processing applications. It included social security number (yes or no) to denote a social security number match or nonmatch, citizenship (yes or no) to signify a match or nonmatch, income (yes or no) to denote an income match or nonmatch result, a Weighted Real Cash Category (WRCC) value between 1.2 and 6 that corresponded to the WRCC

output in the category weighted assets calculation, a WTAC value between 1.6 and 4.8 that corresponded to the Weighted Tangible Asset Category (WTAC) output in the category weighted assets calculation, a Total Weighted Score (TWS) value between 2.8 and 14 to denote the total weighted score calculation result, and eligibility (eligible or noneligible) from the eligibility score process. This table was also linked to the applicant schema via application ID. An unperformed process was denoted as "--".

- *Social security*: This schema table represented the social security schema in Bendex. It contained social security numbers that had been assigned to an individual. Figure 5 displays a sample of these numbers from the prototype.
- *SSA citizenship*: This table represented the citizenship schema in the SSA database. It provided the legal status of social security number holders. It also contained the individual's first name, middle name, last name, date of birth, and legal status date to denote the date green card holders became permanent residents.
- *Income*: This schema contained taxpayers' income information. For simplicity, the following fields were included: gross income, income type (1040, 1040EZ, 1040A), and year. This schema represented the income schema in the state's IRS federal department.
- *Real cash asset*: This table represented the data extraction from KnowX, which lists individuals' real cash source (checking or savings). The cash-on-hand column was supplemented to include applicants' cash-on-hand disclosure from the web application. These records were stored in U.S dollars.
- *Tangible assets*: This table represents the data extracted from KnowX, which lists individuals' tangible asset resources. It included Asset 1 (cars and their market value), Asset 2 (boats and their market value), Asset 3 (houses and their market value), and Asset

4 (watercraft and their market value). The prototype included more assets and asset types than what is presented in Figure 6. The assets' market value was stored in U.S dollars.

The tables were used to identify the type of information necessary for fraud detection in Medicaid applications. However, the schema attributes are just a sample that we conclude in the prototype. Real implementation would include much more information to accurately complete all required information pertaining to the system. In the following section, the type of association and integration between these tables is described.

4.2 Integration & Consolidation

The Medicaid Eligibility Application System (MEAS) prototype interoperates with a set of integrated databases and acts as the data store for the prototype subprocesses and application functions. The following figure illustrates the logical integration of data source tables.

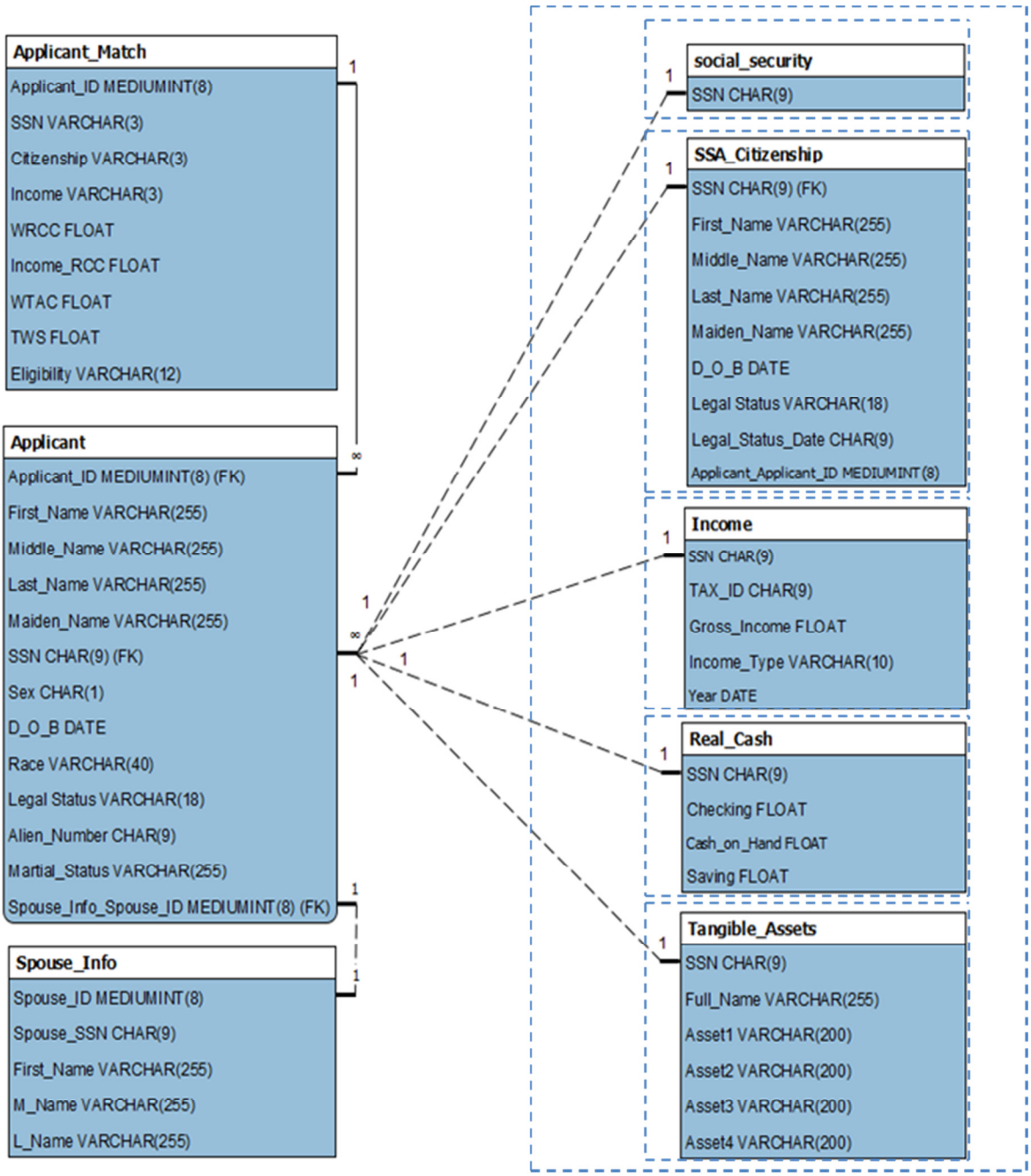

Figure 10 Data-Driven ERD

Figure 10 is an illustration of entity relation diagram (ERD) on data integration between tables. This integration allows data access and data sharing between subprocess applications without the need for an extra layer of integration services. Furthermore, this allows for consistent information that is frequently updated to be readily synchronized for the subprocesses. The

motivation behind data integration was to combine the data located in external sources to present a unified view of these data for the user (Lenzerini, 2002). This development is significantly important when companies merge their databases or when systems combine data results from different internal sources.

The data-driven ERD represents a sample of data integration. A real system may include more tables to completely streamline and synchronize data for applications within the infrastructure. Data consolidation may occur in one or more tables depending on the requirement. For instance, data can be consolidated according to application information in the applicant's application table or according to a number of subprocesses.

CHAPTER 5

Validation

5.1 Data Set

The experimental results were based on inspecting the application layer process for each applicant's Medicaid application, starting with a social security number match and ending with an eligible or noneligible application result. Data populated in the prototype database tables included thousands of synthetic data records.

5.1.1 Synthetic data generator. Data populated in MEAS were created via two applications—Spawner and Generatedata.com—to test the proposed fraud detection mechanisms. Spawner is a win32 application available online (*Spawner Data Generator*, n.d.). The application allows researchers, or anyone for that matter, to generate a random of sample of test data for any type of database through delimited text or SQL insert statements output. It can also output data directly into a MySQL 5.x database. Figure 11 displays a screenshot of fields to be populated or generated with random data based on the assigned parameters required. Spawner is capable of generating as many records as needed. Its capabilities were examined, and its value and performance were noted as it generated over 10,000 social security records.

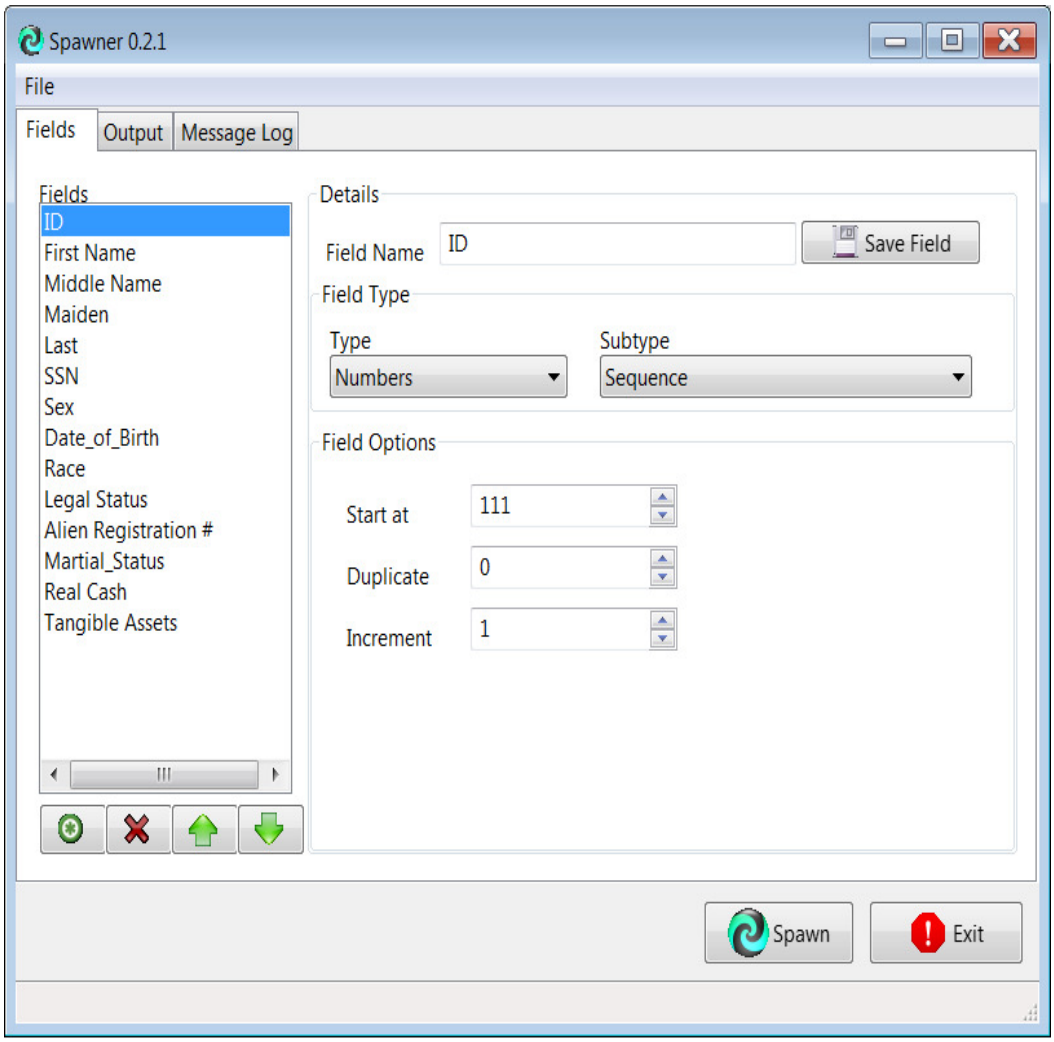

Figure 11 Spawner Data Set Generator

Generatedata.com is another sample/test data generator. It is available via a web form or in a GNU-licensed version that requires a server setup. Nevertheless, it is a free, open-source tool that allows users to generate large volumes of random data in a variety of formats for such purposes as testing software, populating tables in a database, or creating custom data. Alternatively, if you want to avoid setting it up on your own server, you can donate $20 or more to gain a premium account on this site, permitting you to generate up to 5,000 records at a time (instead of the maximum 100), and allows you to save your data sets (*Generatedata.com*, n.d.).

Figure 12 is a screenshot of the spouse info parameter structure before random data were generated for the table.

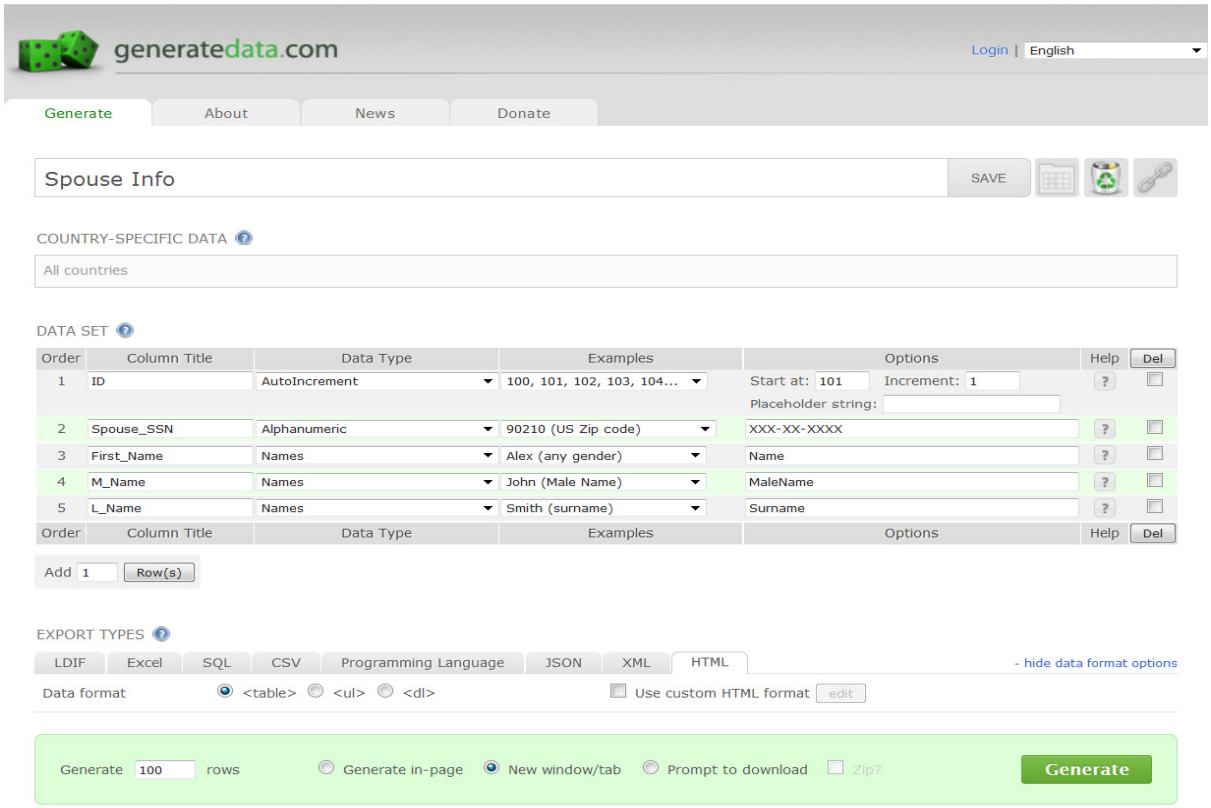

Figure 12 Generatedata.com Data Set Generator

5.2 Synthetic Data Record Validation

This section contains a sample of the data records generated from the aforementioned applications and the experimental results. Each of the following sample applicants was indicated as eligible or noneligible through the data-driven prototype to filter fraudulent Medicaid applications. The operation was followed accordingly to the eligibility flow model described in section 3.3.2.

- **Applicant 101**: Demetria Giselle Clay is represented in Table 4, which displays all of the personal information she has filed in her Medicaid application.

Table 4

Applicant Database Table

ID	First Name	Middle Name	Maiden	Last	SSN	Sex	D.O.B	Race	Legal Status	Alien R. Number	Marital Status
101	Demetria	Giselle	Shelton	Clay	404-15-8841	F	1990/01/27	B	Citizen	-	Married
102	Halee	Martena	Boyd	Sharpe	533-12-7787	F	2001/04/28	W	Non-Citizen	-	Married
103	Natalie	Jonah	Fox	Saunders	589-75-2785	F	1978/11/08	C	Citizen	-	Single
104	Risa	Chelsea	Ramos	Clevelan	646-69-2753	F	1957/12/10	C	P. Resident	493-116-189	Married
105	Conan	Doris		Greene	653-77-6364	M	1987/07/12	W	P. Resident	376-431-477	Single
106	Barrett	Caldwell		Vargas	662-13-0314	M	1979/08/28	B	Non-Citizen	-	Married
107	Ivory	Mannix		Cox	730-56-2993	M	1978/10/02	B	Citizen	-	Divorced
108	Linda	Ferris	Pitts	Herman	742-34-1690	F	1984/11/09	I	Citizen	-	Married
109	Francesca	Walker		Prince	761-75-3769	M	1982/10/21	C	Citizen	-	Widowed
110	Giselle	Imelda	Erickson	Shelton	790-09-4575	F	1979/01/21	P	Citizen	-	Married

Table 5 displays the spouse information Demetria filed with her application. The spouse information record is linked to her application ID.

Table 5

Spouse Information Database Table

ID	Spouse SSN	First Name	M. Name	L. Name
101	021-99-1278	Myra	Tyson	Leilani
102	035-39-3347	Keay	Anne	Donaldson
103	032-91-5618	Audra	Whitney	Sanders
104	041-45-0075	Kathleen	Heidi	Shepherd
105	050-85-2465	Walker	Edan	Allison
106	070-58-5368	Lara	Moran	Chang
107	084-56-8703	Lani	Sparks	Rios
108	096-71-6432	Jayme	Turner	Weber
109	087-47-6591	Hector	Ira	Cabrera
110	062-28-6813	Imelda	Rios	Vaughn

Demetria's social security number, 404-15-8841, exists in the social security table. Table 6 displays a sample of the social security numbers used to check if the applicant had a valid social security number in the system.

Table 6

Social Security Database Table

SSN
404-15-8841
533-12-7787
589-75-2785
646-69-2753
653-77-6364
352-12-0846
184-81-3315
742-34-1690
761-75-3769
790-09-4575

Therefore, the designation "yes" for her social security number assigned to her an applicant match table record.

Her social security number matched her name and citizenship legal status in the SSA citizenship table, which is represented in Table 7. Therefore, the designation "yes" for citizenship assigned to her an applicant match table record.

Table 7

SSA Citizenship Database Table

SSN	First Name	Middle Name	Maiden	Last	D.O.B	Legal Status	L.S Date
404-15-8841	Demetria	Giselle	Shelton	Clay	2003/12/23	Citizen	-
533-12-7787	Halee	Martena	Boyd	Sharpe	2003/12/23	Non-Citizen	-
589-75-2785	Natalie	Jonah	Fox	Saunders	2003/12/23	Citizen	-

Table 7

Cont.

646-69-2753	Risa	Chelsea	Ramos	Clevelan	2003/12/23	P. Resident	2003/12/23
653-77-6364	Conan	Doris		Greene	2003/12/23	P. Resident	2010/12/23
662-13-0314	Barrett	Caldwell		Vargas	2003/12/23	Non-Citizen	-
730-56-2993	Ivory	Mannix		Cox	2003/12/23	Non-Citizen	-
742-34-1690	Linda	Ferris	Pitts	Herman	2003/12/23	Citizen	-
761-75-3769	Francesca	Walker		Prince	2003/12/23	Citizen	-
790-09-4575	Giselle	Imelda	Erickson	Shelton	2003/12/23	Non-Citizen	-

Demetria's gross income, retrieved from income table as illustrated in Table 8, was $9,000.

Table 8

Income Database Table

SSN	Gross Income	Income Type	Year
404-15-8841	$ 9, 000.00	1040 EZ	2013
589-75-2785	$11, 000.00	1040	2013
646-69-2753	$10,000.00	1040 EZ	2013
742-34-1690	$11,000.00	1040 A	2013
761-75-3769	$10, 000.00	1040	2013

This falls below the poverty line. Therefore, an income "yes" match was assigned to her an applicant match record. Demetria's real cash assets were retrieved from the real cash table using her social security number as represented in Table 9. Her assets were $175 ($25 + $100 + $50).

Table 9

Real Cash Database Table

SSN	Checking	Saving	Cash on Hand
404-15-8841	$ 25.00	$ 100.00	$ 50.00
589-75-2785	$ 500.00	$ 950.00	$ 0.00
646-69-2753	$ 5000.00	$ 30,0000.00	$ 100.00
742-34-1690	$ 1000.00	$ 2,500.00	$ 200.00
761-75-3769	$ 150.00	$ 75.00	$ 85.00

This corresponded to a score of 1, as represented in Real Cash Parameters Table (Pg. 31). Therefore, her assigned WRCC value in her applicant match record was 0.60 (1 x 0.60). An income and RCC "yes" match assigned to her an applicant match record since $9,175 ($175.00 +$9,000) is still below the required poverty line. Demetria's tangible assets, retrieved from the tangible asset resources table in Table 10 equaled $453,000 ($8,000 + $8,000 + $435,000 + $2,000).

Table 10

Tangible Asset Database Table

SSN	Asset 1	Asset 2	Asset 3	Asset 4
404-15-8841	Vehicle, $ 8000.00	Boat, $ 8000.00	House, $ 435K	Water Craft, $ 2K
589-75-2785	Vehicle, $ 0.00	Boat, $ 0.00	House, $ 0.00	Water Craft, $ 0.00
646-69-2753	Vehicle, $ 500.00	Boat, $ 0.00	House, $ 0.00	Water Craft, $ 1K
742-34-1690	Vehicle, $ 1000.00	Boat, $ 0.00	House, $ 0.00	Water Craft, $ 0.00
761-75-3769	Vehicle, $ 1000.00	Boat, $ 0.00	House, $ 125K	Water Craft, $ 0.00

Her tangible assets corresponded to a score of 8, as represented in Tangible Asset Parameters Table. Therefore, her assigned WTAC value in her applicant match record was 3.2 (8 x 0.40) as represented in Table 11.

Table 11

Applicant Match Database Table

ID	SSN	Citizenship	Income	WRCC	Income & WRCC	WTAC	TWS	Eligibility
101	Yes	Yes	Yes	Yes	Yes	3.2	3.8	Non-Eligible
102	Yes	No	-	-	-	-	-	Non-Eligible
103	Yes	Yes	Yes	Yes	No	0.8	2.6	Non-Eligible
104	Yes	Yes	Yes	Yes	No	0.8	3.8	Non-Eligible
105	Yes	No	-	-	-	-	-	Non-Eligible
106	No	No	-	-	-	-	-	Non-Eligible
107	No	No	-	-	-	-	-	Non-Eligible
108	Yes	Yes	Yes	Yes	No	0.8	3.8	Non-Eligible
109	Yes	Yes	Yes	Yes	Yes	0.8	1.4	Eligible
110	Yes	No	-	-	-	-	-	Non-Eligible

Demetria's total weighted score was equal to 3.8 (0.60 + 3.2). Because her total weighted score was greater than the threshold, she was identified as noneligible.

- **Applicant 102**: Halee Martena Sharpe's social security number existed in the social security table. Therefore, a social security number "yes" match was assigned to her applicant match schema record. Her social security number matched noncitizen legal status. Therefore, a citizenship "no" match was assigned to her applicant match schema record, and the process skipped to eligibility. Halee's eligibility was identified as noneligible because she did not have a match for citizenship.

- **Applicant 103**: Natalie Jonah Saunders's social security number existed; therefore a social security number "yes" match was assigned. Her social security number matched her name and citizenship legal status; therefore a citizenship "yes" match was assigned. Her gross income was retrieved from the income table and was below the poverty line. Therefore, an income "yes" match was assigned. Her real cash assets were retrieved from the real cash table, and the corresponding score equaled 1.8. An income and RCC "no" match was assigned to her applicant match record since her income and real cash assets were greater than the poverty line. Her tangible assets retrieved from the tangible asset resources table and the corresponding score equaled 0.80. Natalie's total weighted score was equal to 2.6. Although Natalie's total weighted score was below the eligibility threshold, her eligibility was identified as noneligible as her Income and RCC was a "no" match.
- **Applicant 104**: This applicant was similar to Applicant 103. There was a match on social security number, citizenship (more than 5 years of permanent residency), and income. The difference was that Risa Chelsea Clevelan had a TWS above the threshold, which did not affect the application as she was already noneligible due to a "no" match on income and RCC.
- **Applicant 105**: Conan Doris Greene had a social security number match, but no citizenship match because his legal residency was not at least 5 years. Therefore, he was identified as noneligible.
- **Applicants 106 and 107**: Barrett and Ivory did not have a match on social security number and citizenship. Therefore, they were identified as noneligible applicants.

- **Applicant 108**: This applicant was similar to Applicant 104. There was a match on social security number, citizenship, and income. But there was a "no" match on income and RCC. Therefore, Linda Ferris Pitts was identified as noneligible.
- **Applicant 109**: Francesca Walker Prince's social security number had a match. She also had a citizenship match, and her gross income and income and RCC were below the poverty line. Her total weighted score was equal to 1.4, which was below the eligibility threshold. Therefore, her eligibility was identified as eligible.
- **Applicant 110**: Giselle Imelda Shelton's social security number exists in Social Security Table. However, her citizenship returned a noncitizenship legal status. Consequently, a citizenship "no" match assigned to her an applicant match table record. Therefore, her eligibility was identified as noneligible as she has no match for citizenship.

Based on the experimental results of these applicants, it was concluded that for an application to be nonfraudulent, it must satisfy the following requirements: (a) the applicant's social security number exists in the social security table, (b) the applicant's citizenship is labeled as "citizen" or "permanent resident" with 5 years residency, (c) the applicant's income is below required poverty line, (d) the applicant's income and RCC are not greater than the required poverty line, and (e) the applicant's total weighted score is equal to less than the eligibility threshold. Otherwise, the Medicaid application is labeled as fraudulent.

Chapter 6

Data Analysis and Findings

6.1 Overview

This chapter presents the data that were created using a synthetic generator approach and processed in relation to fraud-detection objectives. The fundamental goals that motivated the data analysis were to develop a knowledge base of Medicaid benefits savings when filtering fraud and to determine the validity of the proposed data-driven approach compared to the current approach used for determining fraudulent Medicaid application.

6.2 Description of the Data

The data processed for analysis were based on data results similar to the applicant match table, as presented in Figure 20. All match fields are marked with a "yes," and all nonmatched fields are marked with "no," including the WTAC and TWS columns. The exception is the last column, eligibility, which is marked with either "eligible" or "noneligible" values.

6.3 All Possible Scenarios

This section presents all the possible scenarios or combinations of the processed data. There were two different outputs (yes and no) and seven different categories (social security number, citizenship, income, WRCC, income and WRCC, WTAC, and TWS). The different possibilities were calculated using the data-driven combination formula in Equation 3.

$$\textbf{All Possible Scenarios} = \{(\textbf{Output})^{\textbf{Categories}}\} \tag{3}$$

Equation 3 yields (2^7) a total of 128 possible scenarios that may occur during the processing of generating data. These possible scenarios are presented in the Appendix.

6.5 Analysis Method

This section describes the analysis method used to analyze the data collected. IBM predictive analytics software (SPSS) was used. SPSS is a Windows program used to perform data entry and analysis on large amounts of data. Field (2013) was consulted to learn about the SPSS environment and how to utilize the software to meet the objectives of the study. As a result, descriptive statistics using frequency distribution were considered in order to better understand the data. The following section presents the frequency distribution findings for the data.

6.6 Findings

All 1,010 applicants were processed by means of a current approach and the proposed data-driven approach. Table 12 presents the frequencies that resulted when the data were processed under the current approach. Table 13 presents the frequencies that resulted when the data were processed against the proposed data-driven approach.

Table 12

Eligibility Under Current Medicaid Approach

Criteria	Frequency	Percent
Noneligible Applicants	3620	36.2
Eligible Applicants	6390	63.8
Total	**10010**	**100.0**

Table 12 indicates that 6,390 applications were marked eligible against Medicaid's current approach whereas only 3,620 were marked noneligible. Figure 13 displays a graphical pie chart of the frequency distribution output and shows that the percentage of eligible and noneligible applicants was 36% and 64%, respectively.

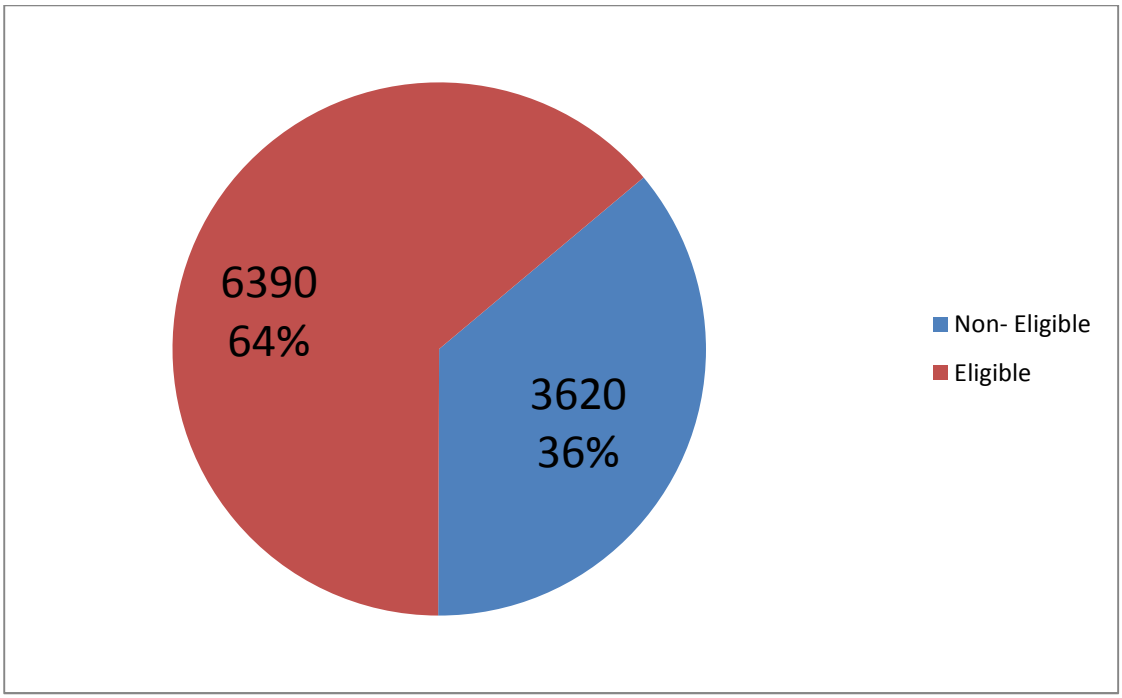

Figure 13 Eligibility Under the current Medicaid approach

Table 5 indicates that 5,776 applications were marked eligible using the data-driven approach whereas only 4,232 were marked noneligible. Figure 14 displays a graphical pie chart of the frequency distribution output and shows that the percentage of eligible and noneligible applicants was 42% and 58%, respectively.

Table 13

Eligibility Under The Data-Driven Approach

Criteria	Frequency	Percent
Non-Eligible Applicants	4234	42.3
Eligible Applicants	5776	57.7
Total	**10010**	**100.0**

.

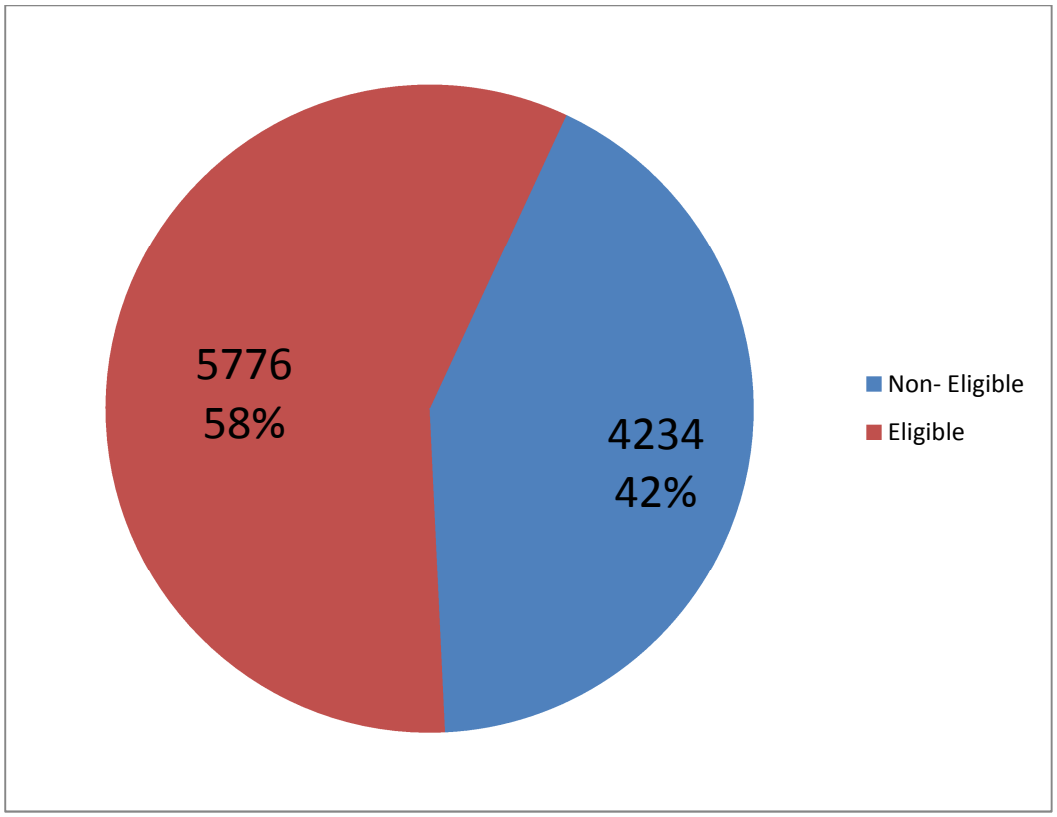

Figure 14 Eligibility Under The Data Driven Approach

Thus, 36.2% of applicants were noneligible based on the current approach compared to 42.3% based on the data-driven approach. This indicates that using the data-driven approach can eliminate more fraudulent Medicaid applications than the current approach used by the Medicaid health services departments. As a result, the frequency descriptive statistics showed that 614 (6,390 – 5,776) more applicants were eligible under the current approach compared to the proposed data-driven approach. Consequently, State's Medicaid services would save $4,250,722 by using the proposed data-driven approach according to Medicaid's average spending ($6,923) per beneficiary (Cassidy, n.d.).

Chapter 7

Conclusion

This thesis discussed fraud-detection ideas within the health care system and alternative approaches for detecting fraud before it occurs. It also provided examples of fraud in the health care system by individuals, facilities, and fraudulent organized entities. It is vital to keep this topic open for continuous study and improvement in order to best utilize federal health care expenditures with minimal or no fraudulent activities allowed.

This thesis also presented related work in many different aspects with regards to health care data mining and fraud-detection tools and techniques. A data-driven implementation that couples a comprehensive, standards-based Medicaid eligibility guideline was proposed with a robust set of fraud-detection workflow processes to filter fraudulent Medicaid applications. Identifying fraud at an early stage reduces the number of abusers of the health care system and allows for future monitoring for similar activities.

The integrated algorithm of weights and scores of asset categories allowed the determination of applicant eligibility based on assets available without undermining each asset value. Furthermore, the synthetic testing data created with the data generator software and processed through IBM SPSS descriptive statistics analysis were examined. As a result, it was determined that state's Medicaid services could use the proposed data-driven system to filter fraudulent Medicaid application and save significant amount of Medicaid expenditures.

References

Agrawal, R., El-Bathy, N., & Seay, C. (2012). Medicaid fraud detection using data broker services. *ACM SIGHIT Record*, *2*(1), 25–25.

Ahsan, K., Shah, H., & Kingston, P. (2010). Healthcare Modelling through Enterprise Architecture: A Hospital Case. In *2010 Seventh International Conference on Information Technology: New Generations (ITNG)* (pp. 460–465). doi:10.1109/ITNG.2010.190

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006). A comparative study for outlier detection techniques in data mining. In *2006 IEEE Conference on Cybernetics and Intelligent Systems* (pp. 1–6). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4017846

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting Review*, *71*(4), 443–465.

Becker, D., Kessler, D., & McClellan, M. (2005). Detecting Medicare abuse. *Journal of Health Economics*, *24*(1), 189–210. doi:10.1016/j.jhealeco.2004.07.002

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, *16*(3), 271–309.

Bentley, P. J. (2000). Evolutionary, my dear Watson. *Investigating Committee-Based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims*, 702–709.

Blue Cross and Blue Shield Association. (n.d.). Understanding Healthcare Fraud. Retrieved from http://www.bcbs.com/report-healthcare-fraud/

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, *17*(3), 235–255. doi:10.1214/ss/1042727940

Bruggemann, A. J., Wijma, B., & Swahnberg, K. (2012). Abuse in health care: A concept analysis. *Scandinavian Journal of Caring Sciences*, *26*(1), 123–132. doi:10.1111/j.1471-6712.2011.00918.x

Cahill, M. H., Lambert, D., Pinheiro, J. C., & Sun, D. X. (2002). Detecting fraud in the real world. In *Handbook of massive data sets* (pp. 911–929). Springer.

Canlas Jr, R. D. (2009). *Data mining in healthcare: Current applications and issues* (Master's thesis). Carnegie Mellon University Australia Adelaide.

Cassidy, A. (n.d.). Health policy Briefs: Per capita caps in medicaid. *Health policy Briefs: Per capita caps in medicaid*. Retrieved from http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=90

Centers for Medicare and Medicaid Services. (2014, February 12). Medicare program - General information. Retrieved from http://cms.gov/Medicare/Medicare-General-Information/MedicareGenInfo/index.html files/1376/index.html

Centers for Medicare and Medicaid Services. (n.d.). Eligibility. Retrieved from http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Eligibility/Eligibility.html

Chan, C.L. & Lan, C. . (2001). A data mining technique combining fuzzy sets theory and Bayesian classifier - An application of auditing the health insurance fee. In *Proceedings of the international Conference on Artificial Intelligence IC-AI' 2001* (pp. 402–408).

Copeland, L. (2011). *Detecting Fraudulent Suppliers of Incontinence Briefs*. Reno: University of Nevada.

Cortes, C., & Pregibon, D. (2001). Signature-based methods for data streams. *Data Mining and Knowledge Discovery*, *5*(3), 167–182.

Dechow, P. M., Ge, W., Larson, C. R., Sloan, R. G., & Investors, B. G. (2007). Predicting material accounting manipulations. *AAA 2007 Financial Accounting and Reporting Section (FARS) [Electronic Version]*, *http://ssrn.com/abstract=997483*.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec*. *Contemporary Accounting Research*, *13*(1), 1–36. doi:10.1111/j.1911-3846.1996.tb00489.x

DePalo, P., & Song, Y.-T. (2012). Healthcare interoperability through enterprise architecture. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*. Retrieved from http://dl.acm.org/citation.cfm?id=2184837

Dionne, G., Giuliano, F., & Picard, P. (2009). Optimal auditing with scoring: Theory and Application to insurance fraud. *Management Science*, *55*(1), 58–70. doi:10.1287/mnsc.1080.0905

Dolins, S. B., & Kero, R. E. (2006). Data management challenges for U.S. healthcare providers. In *Information Resources Management Association; Emerging trends and challenges in information technology management* (p. 724).

Dunn, P. (2004). The impact of insider power on fraudulent financial reporting. *Journal of Management*, *30*(3), 397–412.

El-Sappagh, S. H., El-Masri, S., Riad, A. M., & Elmogy, M. (2013). Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare. *International Journal of Engineering Research and Applications*, *3*, 900–906.

Exodus Payment Systems. (n.d.). The BioPIN-Gateway and the Community Benefit Card: Stopping Fraud Before It Starts. Retrieved from

https://www.exoduspaymentsystems.com/solutions_files/EPS_Medicaid%20WhitePaper.pdf

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1*(3), 291–316.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Federal Bureau of Investigation. (n.d.). Financial-crime-2007. Retrieved from http://www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm#health

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics and sex and drugs and rock "n" roll*. London, UK: Sage.

Generatedata.com. (n.d.). Retrieved from http://www.generatedata.com/

Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. In *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences* (Vol. 3, pp. 621–630). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=323314

Grosser, H., Britos, P., & García-Martínez, R. (2005). Detecting fraud in mobile telephony using neural networks. In *Innovations in Applied Artificial Intelligence* (pp. 613–615). Springer. Retrieved from http://link.springer.com/chapter/10.1007/11504894_85

He, H., Wang, J., Graco, W., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, *13*, 329–336.

Healy, W. L., Ayers, M. E., Iorio, R., Patch, D. A., Appleby, D., & Pfeifer, B. A. (1998). Impact of a clinical pathway and implant standardization on total hip arthroplasty: a clinical and economic study of short-term patient outcome. *Journal of Arthroplasty*, *13*(3), 266–276.

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., & Melton, G. B. (2011). Bias associated with mining electronic health records. *Journal of Biomedical Discovery and Collaboration*, *6*, 48.

Hwang, S.-Y., Wei, C.-P., & Yang, W.-S. (2004). Discovery of temporal patterns from process instances. *Computers in Industry*, *53*, 345–364.

IBM Corporation. (n.d.). IBM Smarter Analytics Signature Solution for healthcare. Retrieved from http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=SP&htmlfid=ZSS03070USEN

Institute for HealthCare Consumerism. (n.d.). LexisNexis Leads Discussion at AHIP Institute 2012 on Leveraging Big Data in Health Care. *Communities: Health Data Analytics LexisNexis Leads Discussion at AHIP Institute 2012 on Leveraging Big Data in Health Care*. Retrieved from http://www.theihcc.com/en/communities/health_care_data_analytics/ lexisnexis-leads-discussion-at-ahip-institute-2012_h3lsfim7.html

Internal Revenue Service. (n.d.). Internal revenue manual (IRM): Communications and liaison. *Internal Revenue Manual (IRM): Communications and Liaison*. Retrieved from http://www.irs.gov/irm/part11/irm_11-004-002.html

Ireson, C. L. (1997). Critical pathways: Effectiveness in achieving patient outcomes. *Journal of Nursing Administration*, *27*(6), 16–23.

Khosrow-Pour, M. (2006). *Emerging trends and challenges in information technology management*. Washington, DC: Mehdi Khosrowpour Idea Group.

King, J., & Malida, J. (n.d.). Before Claims Fraud, What About Eligibility Fraud? Retrieved from http://www.healthcarepayernews.com/sites/healthcarepayernews.com/files/Before%20Claims%20Fraud.pdf

KnowX. (n.d.). KnowX professional. Retrieved from https://knowx.com/

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *19*(2), 65.

Lenzerini, M. (2002). Data integration: A Theoretical Perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 233–246). New York, NY: ACM. doi:10.1145/543613.543644

LexisNexis. (n.d.). LexisNexis Leads Discussion at AHIP Institute 2012 on Leveraging Big Data in Health Care. Retrieved from http://www.theihcc.com/en/communities/health_care_data_analytics/ lexisnexis-leads-discussion-at-ahip-institute-2012_h3lsfim7.html

Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science*, *11*, 275–287. doi:10.1007/s10729-007-9045-4

Mears, B., & Dun & Bradstreet. (2012). Medicaid Fraud Prevention & Detection: Best Practices for combating fraud, waste, and abuse. Retrieved February 18, 2014, from http://www.dnb.com/government/lc/whitepaper-medicaid-fraud-prevention-detection-best-practices-for-combating-fraud-waste-and-abuse.html#.UwKq-rS2HWE

Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8), 44–47.

National Health Care Anti-Fraud Association. (2005). Health care fraud: A serious and costly reality for all Americans. Retrieved from http//www.nhcaa.org/about_health_care_fraud

National Health Care Anti-Fraud Association. (n.d.-a). Consumer Info & Action - What is health care fraud? Retrieved from http://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx

National Health Care Anti-Fraud Association. (n.d.-b). Fighting Health Care Fraud: An Integral Part of Health Care Reform. Retrieved from http://www.nhcaa.org/media/5997/fighting_health_care_fraud_nhcaajune2009.pdf

Nolting, J. (2006). Developing a Neural Network Model for Health Care. In *AMIA Annual Symposium Proceedings*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839654/

North Carolina Department of Health and Human Services. (n.d.-a). Medicaid and Health Choice Applications. Retrieved from http://www.ncdhhs.gov/dma/medicaid/applications.htm

North Carolina Department of Health and Human Services. (n.d.-b). North Carolina Department of Health and Human Services Program Integrity Section. Retrieved from www.ncdhhs.gov/dma/mcac/20120518MCAC-PI-IBMPresentation.pdf

North Carolina Department of Health and Human Services Program Integrity Section. (n.d.). Retrieved from http://webcache.googleusercontent.com/search?q=cache:UYwTp53hXywJ:www.ncdhhs.gov/dma/mcac/20120518MCAC-PI-IBMPresentation.pdf+&cd=2&hl=en&ct=clnk&gl=us

North Carolina Department of Transportation. (n.d.). Division of Motor Vehicles. Retrieved from http://www.ncdot.gov/dmv/

North Carolina Division of Medical Assistance. (n.d.). Medicaid Home. Retrieved from http://www.ncdhhs.gov/dma/medicaid/

Ortega, P. A., Figueroa, C. J., & Ruz, G. A. (2006). A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. *DMIN*, *6*, 26–29.

Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., & Shi, Y. (2006). Application of Clustering Methods to Health Insurance Fraud Detection. In *2006 International Conference on Service Systems and Service Management* (Vol. 1, pp. 116–120). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4114418

Pflaum, B., & Rivers, J. (1990). Employer strategies to combat health care plan fraud. *Benefits Quarterly*, *7*(1), 6–14.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *Computers in Human Behavior*, *28*, 1002–1013.

Shan, Y., Jeacocke, D., Murray, D. W., & Sutinen, A. (2008). Mining medical specialist billing patterns for health service management. In *Proceedings of the 7th Australasian Data Mining Conference* (pp. 105–110). Glenelg, Australia: Australian Computer Society.

Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, *39*, 7441–7450.

SmartyStreets. (n.d.). Verify addresses in real-time. Retrieved from http://smartystreets.com/products/liveaddress-api

Social Security Administration. (n.d.). Data exchanges: Data exchange Programs and Systems. Retrieved from http://www.ssa.gov/gix/programsAndDataExchanges.html

Sokol, L., Garcia, B., West, M., Rodriguez, J., & Johnson, K. (2001). Precursory steps to mining HCFA health care claims. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences* (p. 10). IEEE.

Spawner data generator. (n.d.). Retrieved from http://sourceforge.net/projects/spawner/

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*, *2*, 250–255.

Summers, S. L., & Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis. *Accounting Review*, 131–146.

Tagaris, A., Konnis, G., Benetou, X., Dimakopoulos, T., Kassis, K., Athanasiadis, N., … Koutsouris, D. (2009). Integrated Web Services Platform for the facilitation of fraud detection in health care e-government services. In *9th International Conference on Information Technology and Applications in Biomedicine* (pp. 1–4). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5394355

Thiruvadi, S., & Patel, S. C. (2011). Survey of Data-mining Techniques used in Fraud Detection and Prevention. *Information Technology Journal*, *10*(4), 710–716.

Travaille, P., Müller, R. M., Thornton, D., & Hillegersberg, J. (2011). *Electronic Fraud Detection in the US Medicaid Healthcare Program: Lessons Learned from Other Industries*. Retrieved from http://doc.utwente.nl/78000/

U.S. Government Accountability Office. (2011). Fraud detection systems: Centers for Medicare and Medicaid Services needs to Ensure more widespread use. Retrieved from http://www.gao.gov/products/GAO-11-475

U.S. GAO - Fraud Detection Systems: Centers for Medicare and Medicaid Services Needs to Ensure More Widespread Use. (n.d.). Retrieved from http://www.gao.gov/products/GAO-11-475

Viaene, S., Derrig, R. A., & Dedene, G. (2004). A case study of applying boosting Naive Bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, *16*(5), 612–620.

Walker, L. O., & Avant, K. C. (2005). Strategies for theory construction in nursing. *Pearson/Prentice Hall*.

Wei, C., Hwang, S., & Yang, W.-S. (2000). Mining frequent temporal patterns in process databases. In *Proceedings of 10th International Workshop on Information Technologies and Systems (WITS00)* (pp. 175–180). Brisbane, Australia.

Yang, J. G. S. (2006). Data mining techniques for auditing attest function and fraud detection. *Journal of Forensic & Investigative Accounting*, *1*(1), 4–10.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*, 597–604.

Yang, W.-S., & Hwang, S.-Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*, *31*(1), 56–68. doi:10.1016/j.eswa.2005.09.003

Yueh, F., & Barry, S. (2010). Fraud and Forensics: New techniques, better Results. Retrieved February 18, 2014, from http://webcache.googleusercontent.com/search?q=cache:_zVzkbtpqRIJ:www.nasact.org/conferences_training/nasact/conferences/AnnualConferences/2010AnnualConference/CS14_Barry.pdf+&cd=1&hl=en&ct=clnk&gl=us

Appendix

Table

List of possible scenarios for data-driven processing

	SSN	**Citizen**	**Income**	**WRC**	**WRC & Income**	**WTAC**	**TWS**	**Eligibility**
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes	Yes	Yes	No	No
3	Yes	Yes	Yes	Yes	Yes	No	Yes	No
4	Yes	Yes	Yes	Yes	Yes	No	No	No
5	Yes	Yes	Yes	Yes	No	Yes	Yes	No
6	Yes	Yes	Yes	Yes	No	Yes	No	No
7	Yes	Yes	Yes	Yes	No	No	Yes	No
8	Yes	Yes	Yes	Yes	No	No	No	No
9	Yes	Yes	Yes	No	Yes	Yes	Yes	No
10	Yes	Yes	Yes	No	Yes	Yes	No	No
11	Yes	Yes	Yes	No	Yes	No	Yes	No
12	Yes	Yes	Yes	No	Yes	No	No	No
13	Yes	Yes	Yes	No	No	Yes	Yes	No
14	Yes	Yes	Yes	No	No	Yes	No	No
15	Yes	Yes	Yes	No	No	No	Yes	No
16	Yes	Yes	Yes	No	No	No	No	No
17	Yes	Yes	No	Yes	Yes	Yes	Yes	No
18	Yes	Yes	No	Yes	Yes	Yes	No	No
19	Yes	Yes	No	Yes	Yes	No	Yes	No
20	Yes	Yes	No	Yes	Yes	No	No	No
21	Yes	Yes	No	Yes	No	Yes	Yes	No
22	Yes	Yes	No	Yes	No	Yes	No	No
23	Yes	Yes	No	Yes	No	No	Yes	No
24	Yes	Yes	No	Yes	No	No	No	No
25	Yes	Yes	No	No	Yes	Yes	Yes	No
26	Yes	Yes	No	No	Yes	Yes	No	No
27	Yes	Yes	No	No	Yes	No	Yes	No
28	Yes	Yes	No	No	Yes	No	No	No
29	Yes	Yes	No	No	No	Yes	Yes	No
30	Yes	Yes	No	No	No	Yes	No	No
31	Yes	Yes	No	No	No	No	Yes	No
32	Yes	Yes	No	No	No	No	No	No
33	Yes	No	Yes	Yes	Yes	Yes	Yes	No
34	Yes	No	Yes	Yes	Yes	Yes	No	No
35	Yes	No	Yes	Yes	Yes	No	Yes	No
36	Yes	No	Yes	Yes	Yes	No	No	No
37	Yes	No	Yes	Yes	No	Yes	Yes	No

Table (Cont.)

List of possible scenarios for data-driven processing

	SSN	Citizen	Income	WRC	WRC & Income	WTAC	TWS	Eligibility
38	Yes	No	Yes	Yes	No	Yes	No	No
39	Yes	No	Yes	Yes	No	No	Yes	No
40	Yes	No	Yes	Yes	No	No	No	No
41	Yes	No	Yes	No	Yes	Yes	Yes	No
42	Yes	No	Yes	No	Yes	Yes	No	No
43	Yes	No	Yes	No	Yes	No	Yes	No
44	Yes	No	Yes	No	Yes	No	No	No
45	Yes	No	Yes	No	No	Yes	Yes	No
46	Yes	No	Yes	No	No	Yes	No	No
47	Yes	No	Yes	No	No	No	Yes	No
48	Yes	No	Yes	No	No	No	No	No
49	Yes	No	No	Yes	Yes	Yes	Yes	No
50	Yes	No	No	Yes	Yes	Yes	No	No
51	Yes	No	No	Yes	Yes	No	Yes	No
52	Yes	No	No	Yes	Yes	No	No	No
53	Yes	No	No	Yes	No	Yes	Yes	No
54	Yes	No	No	Yes	No	Yes	No	No
55	Yes	No	No	Yes	No	No	Yes	No
56	Yes	No	No	Yes	No	No	No	No
57	Yes	No	No	No	Yes	Yes	Yes	No
58	Yes	No	No	No	Yes	Yes	No	No
59	Yes	No	No	No	Yes	No	Yes	No
60	Yes	No	No	No	Yes	No	No	No
61	Yes	No	No	No	No	Yes	Yes	No
62	Yes	No	No	No	No	Yes	No	No
63	Yes	No	No	No	No	No	Yes	No
64	Yes	No	No	No	No	No	No	No
65	No	Yes	Yes	Yes	Yes	Yes	Yes	No
66	No	Yes	Yes	Yes	Yes	Yes	No	No
67	No	Yes	Yes	Yes	Yes	No	Yes	No
68	No	Yes	Yes	Yes	Yes	No	No	No
69	No	Yes	Yes	Yes	No	Yes	Yes	No
70	No	Yes	Yes	Yes	No	Yes	No	No
71	No	Yes	Yes	Yes	No	No	Yes	No
72	No	Yes	Yes	Yes	No	No	No	No
73	No	Yes	Yes	No	Yes	Yes	Yes	No
74	No	Yes	Yes	No	Yes	Yes	No	No
75	No	Yes	Yes	No	Yes	No	Yes	No
76	No	Yes	Yes	No	Yes	No	No	No

Table (Cont.)

List of possible scenarios for data-driven processing

	SSN	Citizen	Income	WRC	WRC & Income	WTAC	TWS	Eligibility
77	No	Yes	Yes	No	No	Yes	Yes	No
78	No	Yes	Yes	No	No	Yes	No	No
79	No	Yes	Yes	No	No	No	Yes	No
80	No	Yes	Yes	No	No	No	No	No
81	No	Yes	No	Yes	Yes	Yes	Yes	No
82	No	Yes	No	Yes	Yes	Yes	No	No
83	No	Yes	No	Yes	Yes	No	Yes	No
84	No	Yes	No	Yes	Yes	No	No	No
85	No	Yes	No	Yes	No	Yes	Yes	No
86	No	Yes	No	Yes	No	Yes	No	No
87	No	Yes	No	Yes	No	No	Yes	No
88	No	Yes	No	Yes	No	No	No	No
89	No	Yes	No	No	Yes	Yes	Yes	No
90	No	Yes	No	No	Yes	Yes	No	No
91	No	Yes	No	No	Yes	No	Yes	No
92	No	Yes	No	No	Yes	No	No	No
93	No	Yes	No	No	No	Yes	Yes	No
94	No	Yes	No	No	No	Yes	No	No
95	No	Yes	No	No	No	No	Yes	No
96	No	Yes	No	No	No	No	No	No
97	No	No	Yes	Yes	Yes	Yes	Yes	No
98	No	No	Yes	Yes	Yes	Yes	No	No
99	No	No	Yes	Yes	Yes	No	Yes	No
100	No	No	Yes	Yes	Yes	No	No	No
101	No	No	Yes	Yes	No	Yes	Yes	No
102	No	No	Yes	Yes	No	Yes	No	No
103	No	No	Yes	Yes	No	No	Yes	No
104	No	No	Yes	Yes	No	No	No	No
105	No	No	Yes	No	Yes	Yes	Yes	No
106	No	No	Yes	No	Yes	Yes	No	No
107	No	No	Yes	No	Yes	No	Yes	No
108	No	No	Yes	No	Yes	No	No	No
109	No	No	Yes	No	No	Yes	Yes	No
110	No	No	Yes	No	No	Yes	No	No
111	No	No	Yes	No	No	No	Yes	No
112	No	No	Yes	No	No	No	No	No
113	No	No	No	Yes	Yes	Yes	Yes	No
114	No	No	No	Yes	Yes	Yes	No	No
115	No	No	No	Yes	Yes	No	Yes	No

Table (Cont.)

List of possible scenarios for data-driven processing

	SSN	Citizen	Income	WRC	WRC & Income	WTAC	TWS	Eligibility
116	No	No	No	Yes	Yes	No	No	No
117	No	No	No	Yes	No	Yes	Yes	No
118	No	No	No	Yes	No	Yes	No	No
119	No	No	No	Yes	No	No	Yes	No
120	No	No	No	Yes	No	No	No	No
121	No	No	No	No	Yes	Yes	Yes	No
122	No	No	No	No	Yes	Yes	No	No
123	No	No	No	No	Yes	No	Yes	No
124	No	No	No	No	Yes	No	No	No
125	No	No	No	No	No	Yes	Yes	No
126	No	No	No	No	No	Yes	No	No
127	No	No	No	No	No	No	Yes	No
128	No	No	No	No	No	No	No	No